Proceedings of the 10th International Conference on Preservation of Digital Objects

iPRES2013

3-5 SEPTEMBER 2013
LISBON - PORTUGAL

Editors: José Borbinha, Michael Nelson, Steve Knight

TÉCNICO LISBOA

inesc id lisboa

Blank page

iPRES 2013 – 10th International Conference on Preservation of Digital Objects

Biblioteca Nacional de Portugal
Campo Grande, 83
1749-081 Lisboa
Portugal

Cover Design: João Edmundo

ISBN 978-972-565-493-4

Blank page

PREFACE

From September 3-5, 2013 (with tutorials the September 2 and workshops lasting until the September 6), the Instituto Superior Técnico was pleased to host the tenth annual iPRES Conference in cooperation with DC-2013, the International Conference on Dublin Core and Metadata Applications. Previous iPRES conferences were held in Beijing (2004, 2007), Göttingen (2005), Ithaca, NY (2006), London (2008), San Francisco (2009), Vienna (2010), Singapore (2011), and Toronto (2012). The next conferences are planned for Melbourne (2014), and Chapel Hill (2015).

The Organizing Committee was pleased to note that the event continued to garner significant interest, with well over 60 submissions received from 21 countries around the world, with most proposals coming from Europe and North America. In conjunction with DC-2013, nine workshops and ten tutorial sessions were accepted, as well as 16 full papers, 20 short papers, and two panel presentations delivered during 16 sessions.

The conference hosted three keynote presentations. On the first day, Gildas Illien of the BNF presented "Darling, We Need to Talk". On the second day, Paul Bertone, from the European Bioinformatics Institute, spoke on the potential of "Digital information storage in DNA". Finally, on the final day, Carlos Morais Pires, from the European Commission, presented "Data Infrastructures in Horizon2020: support to data and computing intensive science". Technical sessions at the conference were on central preservation topics like web archives, digital object preservation, sharing knowledge, repositories, preservation in the corporate world, cooperation, data and beyond data, national strategies, assessment, technologies, governance, collections, and curation.

The conference also hosted an exciting poster and demo session that showcased excellent early results and software demonstrations. The poster and demo session, along with the reception that followed, proved to be an excellent opportunity for academics, students, industry representatives and other professionals involved in digital preservation to network and share information.

Several corporate and academic sponsors generously assisted the work of iPRES 2013: ExLibris, OCLC, Tessella, Oracle, Marka, BNP, DANS, and the University of Toronto iSchool. Finally, a team of volunteers of junior researchers from the INESC-ID and MsC and PhD students from the IST contributed to the success of the conference.

The organizing committee was delighted with the success of the conference, and wishes to note that the conference would not have occurred without the efforts of the many members of the program review committee, who gave generously of their time. The program and conference co-chairs also wish to express their gratitude to the local organizers who did so much to make the conference a success and to create a welcoming environment for attendees.

José Borbinha, Conference Committee Chair
Michael L. Nelson, Program Co-Chair
Steve Knight, Program Co-Chair

CONFERENCE ORGANIZATION

General Chair
José Borbinha (IST/INESC-ID, Portugal)

Program Chairs
Michael Nelson (Old Dominion University, United States)
Steve Knight (National Library of New Zealand, New Zealand)

Publicity Chairs
Angela Dappert (Digital Preservation Coalition, United Kingdon)
Jane Greenberg (University of North Carolina at Chapel Hill, United States)

Workshop Chairs
Daniel Gomes (FCCN, Portugal)
Shigeo Sugimoto (University of Tsukuba, Japan)

Posters and Demos Chairs
Christoph Becker (Vienna University of Technology, Austria)
Miguel Ferreira (KEEP SOLUTIONS, Portugal)

Tutorials Chairs
Ana Baptista (University of Minho, Portugal)
Barbara Signori (Swiss National Library, Switzerland)

Programme Committe
Reinhard Altenhoener (National Library of Germany, Germany)
Bjarne Andersen (Netarchive.dk, Denmark)
Gonçalo Antunes (IST/INESC-ID, Portugal)
Andreas Aschenbrenner (Göttingen University, Germany)
Tom Baker (Dublin Core Metadata Initiative, United States)
José Barateiro (LNEC, Portugal)
Christoph Becker (Vienna University of Technology, Austria)
José Borbinha (IST/INESC-ID, Portugal)
Raju Buddharaju (National Library Board, Singapore)
Gerhard Budin (University of Vienna, Austria)
Artur Caetano (IST/INESC-ID, Portugal)
Paul Conway (University of Michigan, United States)
Robin Dale (Lyrasis, United States)
Angela Dappert (Digital Preservatipn Coalition, United Kingdon)
Joy Davidson (Digital Curation Centre, United Kingdon)
Michael Day (UKOLN, University of Bath, United Kingdon)
Janet Delve (University of Portsmouth, United Kingdon)
Angela Di Iorio (Fondazione Rinascimento Digitale, Italy)
Jon Dunn (Indiana University, United States)
Miguel Ferreira (KEEP SOLUTIONS, Portugal)
Kevin Glick (Yale Universit, United States)
Andrea Goethals (Harvard University, United States)
Daniel Gomes (FCCN, Portugal)

Mariella Guercio (University of Rome Sapienza, Digilab, Italy)
Mark Guttenbrunner (Vienna University of Technology, Austria)
Carolyn Hank (McGill University, Canada)
Ross Harvey (Simmons College, United States)
Adam Jatowt (University of Kyoto, Japan)
Leslie Johnston (Library of Congress, United States)
Max Kaiser (Austrian National Library, Austria)
Christopher Khoo (Nanyang Technological University, Singapore)
Ross King (Austrian Institute of Technology, Austria)
Amy Kirchhoff (ITHAKA, United States)
Steve Knight (National Library of New Zealand, New Zealand)
Hannes Kulovits (Vienna University of Technology, Austria)
Cal Lee (University of North Carolina at Chapel Hill, United States)
William Lefurgy (Library of Congress, United States)
Jens Ludwig (Göttingen State and University Library, Germany)
Maurizio Lunghi (Fondazione Rinascimento Digitale, Italy)
Peter May (The British Library, United Kingdon)
Nancy McGovern (Inter-university Consortium for Political and Social Research, United States)
Andrew McHugh (Digital Curation Centre, United Kingdon)
Carlo Meghini (CNR ISTI, Italy)
Salvatore Mele (CERN, Switzerland)
Eva Méndez (Universidad Carlos III, Spain)
Ethan Miller (University of California, Santa Cruz, United States)
David Minor (UC San Diego, United States)
Reagan Moore (University of North Carolina at Chapel Hill, United States)
Michael Nelson (Old Dominion University, United States)
Quyen Nguyen (NARA, United States)
Achim Osswald (Cologne University of Applied Sciences, Germany)
Natalie Pang (Nanyang Technological University, Singapore)
Christos Papatheodorou (Ionian University, Greece)
David Pearson (National Library of Australia, Australia)
Maureen Pennock (The British Library, United Kingdon)
Meg Phillips (NARA, United States)
Diogo Proença (IST/INESC-ID, Portugal)
Andreas Rauber (Vienna University of Technology, Austria)
Cristina Ribeiro (University of Porto, Portugal)
Seamus Ross (University of Toronto, Canada)
Raivo Ruusalepp (Estonian Business Archives, Estonia)
Michael Seadle (Humboldt-Universität zu Berlin, Germany)
Robert Sharpe (Tessella, United Kingdon)
Barbara Sierman (National Library of the Netherlands, Netherlands)
Barbara Signori (Swiss National Library , Switzerland)
Tobias Steinke (National Library of Germany, Germany)
Stefan Strathmann (Göttingen State and University Library, Germany)
Stephan Strodl (SBA, Austria)
Shigeo Sugimoto (University of Tsukuba, Japan)
David Tarrant (University of Southampton, United Kingdon)
Manfred Thaller (Universität zu Köln, Germany)
Emma Tonkin (UKOLN, University of Bath, United Kingdon)
Ilias Trochidis (Tero, Greece)

Bram van der Werf (Open Planets Foundation, Netherlands)
Raymond van Diessen (IBM, Netherlands)
Ricardo Vieira (IST/INESC-ID, Portugal)
Richard Wright (BBC, United Kingdon)

Extra Reviewers
Ahmed Alsum
Elisabeth Weigl
Hany Salaheldeen
Johannes Binder
Stefan Pröll
Tomasz Miksa
Yasmin Alnoamany

Doctoral Symposium Committee
Artur Caetano (IST/INESC-ID, Portugal)
Eva Méndez (Universidad Carlos III, Spain)
Andreas Rauber (SBA, Austria)
Cristina Ribeiro (University of Porto, Portugal)
Adam Jatowt (University of Kyoto, Japan)
Jane Greenberg (University of North Carolina at Chapel Hill, United States)
Gabriel David (University of Porto, Portugal)
Elsa Cardoso (ISCTE-IUL/INESC-ID, Portugal)

Local Committee
José Barateiro (LNEC, Portugal)
Mário Silva (IST/INESC-ID, Portugal)
Gonçalo Antunes (IST/INESC-ID, Portugal)
Ricardo Vieira (IST/INESC-ID, Portugal)
Diogo Proença (IST/INESC-ID, Portugal)
Gilberto Pedrosa (IST/INESC-ID, Portugal)
João Edmundo (IST/INESC-ID, Portugal)
António Higgs (IST/INESC-ID, Portugal)
Marzi Bakhshandeh (IST/INESC-ID, Portugal)
Paulo Ferreira (IST/INESC-ID, Portugal)
Luis Veiga (IST/INESC-ID, Portugal)
Elsa Cardoso (ISCTE-IUL/INESC-ID, Portugal)

CONTENTS

Note: full papers have the title in bold

Studies on the scalability of web preservation

Rory Blevins
Tessella
26, The Quadrant,
Abingdon Science Park
Oxfordshire
OX14 3YS UK
Rory.Blevins@tessella.com

Ismail Patel
Tessella
Chadwick House,
Birchwood Park,
Warrington,
WA3 6AE UK
Ismail.Patel@tessella.com

Jack O'Sullivan
Tessella
26, The Quadrant,
Abingdon Science Park
Oxfordshire
OX14 3YS UK
Jack.O'Sullivan@tessella.com

Ashley Hunter
Tessella
Chadwick House,
Birchwood Park,
Warrington,
WA3 6AE UK
Ashley.Hunter@tessella.com

Robert Sharpe
Tessella
26, The Quadrant,
Abingdon Science Park
Oxfordshire
OX14 3YS UK
Robert.Sharpe@tessella.com

Pauline Sinclair
Tessella
26, The Quadrant
Abingdon Science Park
Oxfordshire
OX14 3YS UK
Pauline.Sinclair@tessella.com

ABSTRACT
This paper describes a mechanism for improving the scalability of preservation actions on large linked archives, such as WARC and ARC files produced from the archiving of web sites.

To enable accurate but efficient preservation actions, information on the files embedded within a container object, such as the file formats of the embedded files, are aggregated and recorded as properties of the container object. This occurs during the ingest of objects into the archiving system, specifically at the characterization stage when files are identified and validated. To ensure that the details of all embedded files are also recorded, nested archives are recursively unpacked and their contents characterized to identify all files in a package. Information about the embedded files is then stored as properties of the container object: this allows us to efficiently aggregate information about the contents of a container as queryable properties of the container.

This storage of the embedded file type information on the container object reduces the number of objects and properties which have to be queried to perform a preservation action, such as migration to a more recent file type. The database can be queried for a specific file type, and all files of that type, and archives containing files of that type will be returned without needing to query each embedded object individually.

Archives containing files in need of preservation are temporarily unpacked and the files in need of transformation identified and migrated. Following the preservation action, the internal links within the archive are updated to maintain the integrity of the archive and the modified objects are re-ingested back into the system.

This approach results in minimal extra overhead at the ingest stage of preservation, but substantially reduces the number of entities which need to be queried to identify objects at risk when performing preservation actions. In the case of large web archives, this may be several orders of magnitude, producing a corresponding increase in performance and scalability.

KEYWORDS
Scalability, Web Archiving, Characterization

1. INTRODUCTION
Many organizations now regularly perform large scale web crawls [Pennock][1]. For example, Bibliothèque nationale de France (BnF) have been performing large scale archiving of web sites since 2002 and by 2011 had accumulated approximately 200TB held within 1.5 million ARC files [2]. These crawls have been managed using web crawling software: initially HTTrack [3], then Heritrix [4] and finally adding the NetarchiveSuite [5], developed by the Royal Library of Copenhagen and the University Library of Aarhus. As can be seen from the size of the accumulated collections, the actual process of collecting web sites can be performed in a reasonable time frame and thus already scales fairly well.

However, typically, such web crawls are not as well characterized as other digital material being ingested into an archival repository. The normal method of such characterization can vary but would typically involve [6]:

- Identification of the format of each file
- Validation of the format of each file
- Property extraction from each file
- Embedded object extraction from each file
- Recursive characterization of such embedded objects using the steps above.

The first three steps are relatively simple and straightforward since these crawls produce a container file in either ARC [7] or WARC [8] format with well-defined properties.

It is also relatively easy to extract the embedded objects from such a container (to produce the original files that manifested the sites crawled) and to characterize each file in turn.

This can produce a very large number of entities to be characterized. For example, the 200TB in the BnF collections are estimated to contain 50 billion embedded objects [9].

Each of these embedded objects can then be assessed to see if they can be adequately preserved in the long-term. This can be done, for example, by comparing the properties (e.g., format of the files) against known issues and then migrating problem files to a new format [6]. This has previously been performed on a small experimental scale [10] but not yet, to the best of the authors' knowledge, on a larger scale.

In fact, it is quite controversial whether or not such format obsolescence exists [11]. A recent study of web material [12] has shown that while most formats have persisted for a decade or more, not all do so and older versions of formats fade from popularity. In a sense, this argument is not relevant to this paper anyway since it is mainly focused on the scalability of bulk operations on large web archives collections and migration is just an example of such operations.

The size of the problem places at least two scalability demands on the ability to preserve websites after crawling that are addressed in this paper:

- Ability to characterize such a large number of files
- Ability to use such properties to determine future migration strategies.

This is caused by both the amount of computing resources needed to characterize, say, 50 billion entities and the ability to cope with the amount of information that such characterization produces and still make it useful in future preservation actions in a timely manner.

The latter issue (i.e. coping with this amount of information) occurs because it will strain the ability of any indexing system to enable searches to be made that can return information in reasonable timescales. Given that the quantity of material on the web is still rapidly increasing and future scans are likely to be more frequent and more comprehensive, this problem is likely to become more pronounced over time, probably outpacing improvements in indexing capabilities.

In this paper we describe an approach whereby we break the problem down into two parts to remove this indexing issue:

- Describing properties at the container level in sufficient detail to determine whether the container requires some action to be applied to it or its content.
- Dealing with each container (and its contents) in turn.

This approach is described in more detail in section 3, and the impact on the characterization process is described in section 4. Section 5 describes the impact on preservation actions (using format migration as an example). Finally, some general conclusions are drawn.

2. METHOD

This work has been carried out using Tessella's Safety Deposit Box (SDB) software. This has an existing suite of web crawling and web characterization functionality that enabled the specific problems to be addressed efficiently. The testing in this paper used version 4.3 of this software.

Performance testing was carried out using an Amazon EC2 M1 medium instance; a single-core Linux instance with the approximate processing power equivalent of two 1.2 GHz Opteron Processors [13].

SDB is commercial software, however most of the tools described in this paper are open-source and the methodology described in this paper should be generally applicable to the preservation of web archives irrespective of the underlying software used to implement the digital preservation repository.

3. CONTAINER VS EMBEDDED OBJECT PROPERTIES

ARC files were developed by the Internet Archive to enable efficient storage of data from web crawls and other archives of website data. WARC files are an ISO-certified extension of this format which allows recording of additional information, such as HTTP request headers and additional metadata (including file conversion records, which hold metadata on files which have been converted into a different format). They have the same basic format: a header block, followed by a series of URL records, which may themselves be compressed with a compression algorithm such as gzip. While they are efficient at storing the results of web crawls, accessing their embedded files requires temporarily unpacking the objects, which can be computationally expensive in the case of large archives.

To correctly preserve the objects embedded in the archive files, a preservation system must be able to identify and characterize both the archive container object, and the objects contained within the archive. It must later be able to query key properties of the embedded objects to determine if they are in need of preservation actions, such as file format migration.

Standard archiving and storage systems can either ignore the contents of container formats, or attempt to record properties of all embedded objects. If a system does not hold information about the individual objects embedded in archive files, any attempt to preserve the archived files, such as migrating them to a newer file format, will either ignore the embedded files completely or else require the system to extract all files from all stored archives to determine which files need preservation actions to be applied. Alternatively, if a system stores a complete set of technical metadata information on each embedded object in the database, accurate searching is possible, but this can result in a very large number of entities which cannot be queried within a reasonable timeframe.

In this study, to reduce the number of objects which have to be queried to locate files at risk, we associate queryable information

about the embedded objects as properties of the container object. In the case of file format migration these properties would be the formats of the files embedded in an entity, each file format being stored as a separate property of the container object with a name such as "this container contains objects of type", with the value being one of the file formats embedded in the object.

This approach requires two modifications to standard workflows: firstly, the characterization process must correctly characterize embedded files and associate the required properties of these files with the container. As archives may themselves contain other archives, this process must be recursive, and characterize all files in all archives contained in a particular object.

Secondly, the preservation process must search for embedded objects in need of preservation by checking the properties of each container object to determine if it contains objects in need of preservation, and then unpack and process those archives which do contain objects in need of preservation.

Because this approach requires that only each container is queried, and not every embedded object, considerable reductions in the number of entities which need to be queried can be achieved. In the case of web archives containing many hundreds or thousands of objects, which is not uncommon in the case of archived web crawls, this can result in a reduction of several orders of magnitude in the number of entities queried, producing a concomitant increase in the speed of identifying objects in need of preservation.

4. CHARACTERIZATION OF WARC FILES

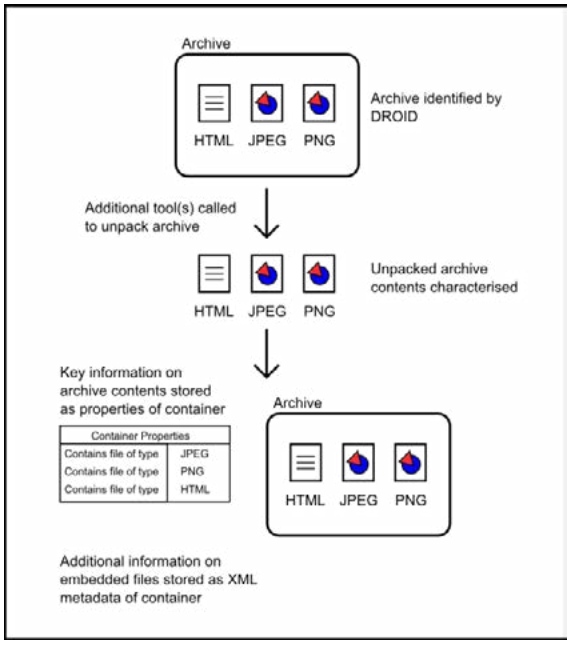


Figure 1.

Schematic of basic steps to aggregate information on archive contents

4.1 Ingest

To correctly preserve web archives, it is necessary to ingest them into the archiving system. This requires a number of steps which are usually automated via a workflow to allow the efficient ingest of objects. In the case of a web archive, the steps required for ingest will typically involve:

- Crawling a given URL and creating a submission information package (SIP) for ingest into the archiving system
- Checking the produced SIP for viruses
- Checking the integrity of the produced SIP: for example, that the number of files in the SIP matches the number of files in the associated technical metadata
- Characterization of the files in the SIP, as described below.
- Storage of the physical files on storage systems
- Storage of the SIP metadata in the repository database

Simultaneously with the ingest, the system also generates XML metadata which describes the ingested objects, their relationships and key properties (e.g. Significant and/or Transformational Information Properties). This metadata is also stored in the database on successful ingest.

4.2 Characterization of files

Characterization is one of the key processes in ingest, and typically involves three steps: identification, validation and property extraction.

To characterize the files embedded in WARC or ARC files, the system must first identify the archive file. SDB identifies files using the open-source DROID (Digital Record Object IDentification) tool, originally developed for The National Archives [14]. DROID identifies files from their byte sequences by searching for signatures specific to file types. The current version of DROID is capable of detecting over one thousand different file types, and its signature definitions are continually updated to improve DROID's capabilities. If files are not identified by DROID, additional tools may also be called.

A file whose format has been identified with DROID will have its technical metadata updated to associate it with a specific Persistent Unique Identifier (PUID) as defined in the PRONOM technical registry, a publically available registry of technical information on file formats. As well as PUIDs, PRONOM and other technical registries provide technical information for the preservation of different file formats. For example, this can include software tools for validation, extraction of key properties or extraction of embedded objects for specific file formats. They may also provide information on tools and pathways for migration between different file formats.

SDB incorporates the data from PRONOM in its own technical registry, which it uses to determine the appropriate tools for characterizing and migrating each of the different file formats.

Following identification, additional tools will be called to validate file formats and to extract key properties of files to ensure accurate long term preservation of the object can take place within the managed digital repository. Information from

each of these steps is written to the metadata element which represents each file.

4.3 Extraction of Embedded files

After a file has been identified, characterized and undergone property extraction, the registry is checked to determine whether object extraction tools exist for each object in the SIP. Once an object extraction tool has been identified for an archive file format, such as a (W)ARC or ZIP files, the tool is called to extract the contents of the container into a temporary work area. Files extracted from the WARCs are passed to DROID for identification. DROID will attempt to uniquely identify the file format of the file and, if a file is identified, may pass it on to further tools (determined by querying the technical registry) which will validate the file format and extract properties of the object to be maintained as technical metadata within the repository system.

This information is stored as part of the XML metadata of the container object, allowing information on the objects inside a container to be retrieved without recharacterizing the entire contents of the web archive container file.

As discussed in section 3, to enable efficient preservation actions, key properties of the embedded files are aggregated and stored as properties of the container object. For example, the file types of embedded files are stored as separate individual properties of the container file for use in migration and other archival operations that require the efficient identification of archives containing specific file formats.

This process of extraction and characterization is recursive: if archive files are found in the unpacked archive, these too are unpacked and their contents are in turn characterized. Key properties are recorded for embedded archives, both in the metadata entry for the embedded archive, and as part of the properties added to the parent container, so that the original archive file has properties which aggregate information from all levels of embedded file within the archive. For example, in the case of recording file format information for file format migration, the top-level container will have properties, including transformational information properties, representing each of the file formats embedded in any of the contained files. In addition there will be entries in the XML metadata for each embedded file, and in the case of nested archives, this metadata will include properties representing each of the file formats embedded in the nested archive.

4.4 Conceptual Characterization

In addition to the above physical characterization, web sites pass through a conceptual characterization process. This identifies the existence of technology-independent information objects (e.g., a web page, an image or a document) that can be manifested in a variety of different technologies (each potentially changing the number and arrangement of files as well as file formats). This allows the identification of links between these information objects (e.g., the link between a web page and an image). This is then subsequently used to identify information objects that might need modification even though it had not been directly affected by a preservation action (e.g., the need to edit a HTML page if an image has changed format and thus extension).

4.5 Performance of Characterization

To measure the performance of the characterization step on typical web archives, we performed some basic benchmarking to test whether the characterization step was sufficiently fast for efficient web archiving. This involved running SDB 4.3 on the single core cloud computing instance described above, measuring the performance during ingest of a selection of public websites.

The first thing to note is how the relative speed of characterization compared to other ingest steps. Analyzing the time spent on each ingest step (Figure 2) clearly shows that the dominant steps are:

- Crawling (28%)
- Thumbnail creation (58%)
- Characterization (8%)

Web crawling is known to be a limiting step, but the elapsed time is largely dependent on wait times and bandwidth issues.

Creating thumbnails is very process intensive since SDB creates an image of all archived HTML files as they originally appeared, complete with any embedded objects. If increased throughput is required the thumbnail creation step can be disabled, which would result in a considerable improvement in the overall rate of ingest.

Characterization took just 8% of the time of a typical ingest equating to a typical speed of 60MB/min. This is a considerable

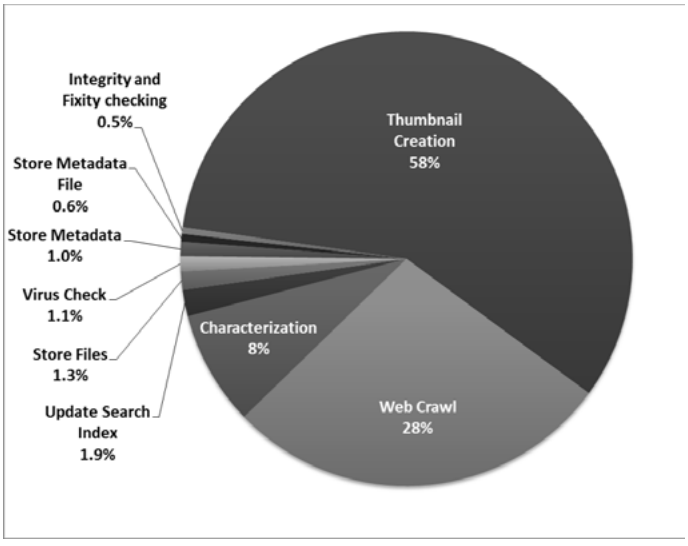


Figure 2

Typical division of time to process ingest steps, as a proportion of total ingest time

improvement on the results reported by the State and university of Denmark using FITS [15] which averaged below 4MB/min. We are not sure of the reason for the discrepancy. One possible reason is that, while SDB will only use a single tool for each characterization process, FITS can attempt to use several. Another possibility is the way jobs are packaged within the workflows of each system could be different.

Figure 3 shows the breakdown of the percentage of the time taken to perform various tasks within the characterization process. This shows that decompressing the WARC files is the most time-consuming of the tasks, taking 51% of the time taken for characterization. It took a total of just 4% of the time for DROID to identify the WARC files and a further 5% for DROID to identify the embedded files post extraction. Jhove and other tools (e.g., SDB's built-in XML validator) took most of the rest of the time to validate and extract file properties (31%). Conceptual characterization took the remaining 9% of the time.

All other ingest processes (i.e. excluding crawling, thumbnail creation and characterization) took just 6% of the time. This includes creating a SIP from the crawl, performing initial quality control checks (for SIP integrity, fixity and virus checks) plus the overhead in storing the resulting content files in a file store, storing the metadata in both a database and a file store and updating a SOLR search index.

One thing that is clear is the extra process of aggregating embedded object properties still allows efficient characterization and ingest of large archives.

4.6 Scaling up

This study did not have access to significant hardware, only using a single-core medium size amazon EC2 cloud instance for benchmarking, which is considerably underpowered compared to the multi-server setups used in many modern web archiving systems. This makes it hard to quantify exactly how this approach would scale if deployed on more significant hardware.

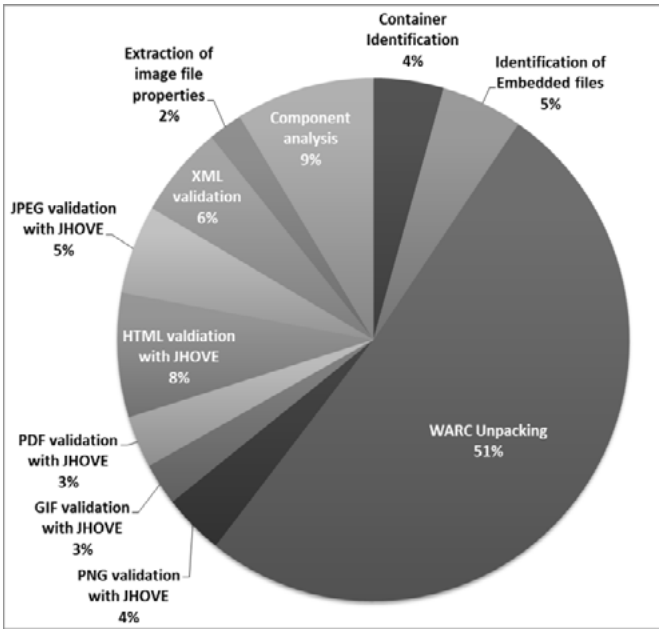


Figure 3

Proportion of characterisation time taken by individual tools

However, it is known that the approach used in this study has been used to achieve total ingest (including characterization) rates in excess of 20TB/day by FamilySearch using relatively modest processing power (2 Dell 2950 servers each with 2 Intel Xeon E5430 processors, with 4 cores clocked at 2.66GHz and 32GB RAM) [16]. Even though this has modest cost for a production system (c. $10k at today's prices) it is many times the processing power of the single core benchmarking instance used in this study. Also, unlike in this study, it enables ingest (and particularly characterization) of different content to be run dozens of time in parallel. Hence, while that study related to much larger files (typically 10-30MB) it did find that the fundamental limit on scalability was the ability to read files fast enough from disk and transfer them across the network and not the processing speed of the server. This required the use of high performance switches and drive arrays to reduce this bottleneck.

While processing more files is likely to lead to a higher overhead, it is still reasonable to expect the method proposed in this study will also parallelize well by adding more computing cores and more server machines.

5. MIGRATION INSIDE WARC FILES

To take advantage of the stored embedded object properties, preservation actions, such as file format migration, must use these properties to reduce the number of entities which must be queried to determine which entities require action.

As with ingest, file format migration in an archiving system normally requires a number of steps to occur through an automated workflow, although a number of key steps can also require human intervention:

- File formats at risk are selected, either by manually selecting a list of PUIDs to migrate or by choosing a risk threshold above which to migrate files
- Files to migrate are chosen from the files at risk
- Pathways to migrate the files at risk are chosen
- The files are migrated, as described below
- The SIP is re-ingested into the database, in a similar manner to the ingest workflow.

5.1 General approach

To migrate files inside (W)ARC containers we used the approach of breaking the problem down into parts:

- Finding which of the millions of containers are in need of a preservation action to be applied to them
- Determining which entities within each container then need action, and extracting these from the container to a temporary working area of the system.
- Performing that action (in this paper a migration will be used as an example of such an action)
- Re-wrapping the content into a new container

This aggregation of the file types of the embedded objects as properties of the container results in a dramatically smaller number of objects to query during preservation actions. This in turn results in a significant improvement in the scalability of performance related to the preservation of web archives and other container formats.

5.2 Finding containers in need of action

In the case of file format migration, file formats at risk are selected either by manually choosing specific file formats which are at risk, or by selecting a risk threshold, a value which indicates how at risk of obsolescence a file is. To determine which file formats are at risk using a risk threshold value, the archiving system must query a technical registry to retrieve file formats which are above this risk threshold. The risk threshold for each file format is determined by answering a number of risk questions about each file format in the technical registry, such as whether it is an open-source or proprietary format. The responses to these question are then weighted (weightings for each question are set by the system user, depending on their requirements) and combined to create a risk value.

Either method produces a list of file formats which require migration, identified by their PUIDs. To locate files in the repository requiring preservation we then query the database for:

- Files of a file format type at risk
- Files which have a property "contains file(s) of type" matching one of the formats at risk.

As technical metadata about these file objects are stored in the system's database, these can be queried easily with a SQL query on the relevant database tables. The resulting list of files and containers at risk is passed onto the next step of the migration process: determining which individual files require migration.

5.3 Determine at risk content within a container

Once an archive has been identified as requiring migration, all files within it are extracted into a temporary work area. As with characterization, this is a recursive process, as archive files may in turn contain further archives. All archive files in a particular archive are in turn unpacked. From this temporary unpacked copy of the archive, the files in need of migration need to be identified.

The properties on a container only indicate that an archive contains a particular file format at risk and not which files within the archive are of that format. This means that once a container containing files at risk has been determined, individual files at risk within the container must be identified. This is achieved by parsing the XML metadata associated with the container for elements representing embedded objects of the file format at risk, or embedded elements which also have a property indicating that they contain files at risk. This occurs recursively, to identify all files at risk even in multiple nested archives.

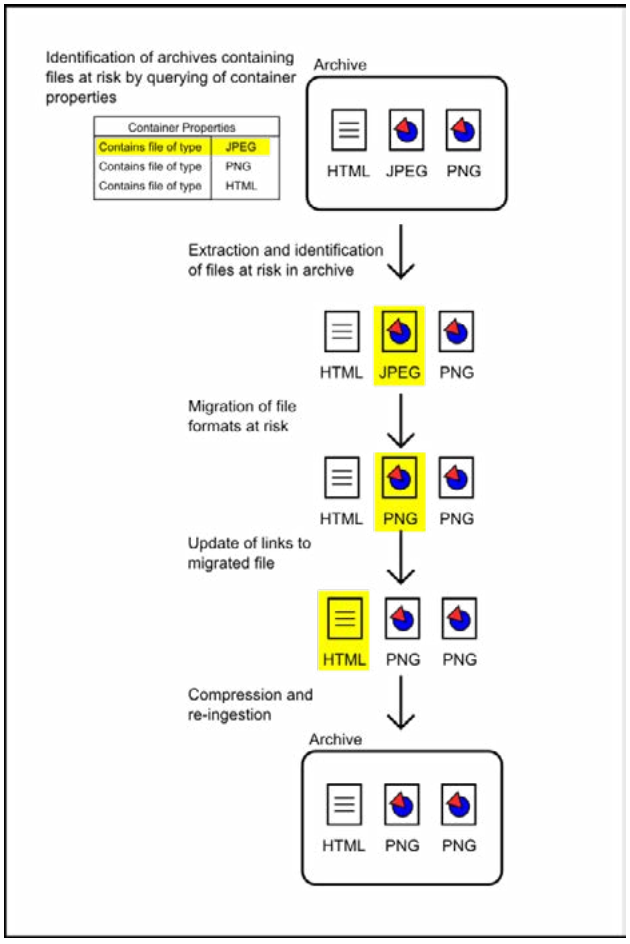


Figure 4

Schematic of basic steps required to perform file format migration on embedded files. In this example, a JPEG to PNG migration is performed.

5.4 Performing migration

Once files at risk have been identified, the user chooses a pathway to transform files at risk into more current formats. Possible migration pathways for a file format at risk are obtained from the technical registry, and one is manually chosen for each format at risk by the user.

The exact pathway and tool used for migration depends on the pathways available in the technical registry, but the general procedure is simple: the tool is invoked, either through its API or through the command line, to convert the file into the new file format in a temporary work area.

Migration requires that not only are particular file formats migrated successfully, but that the conceptual components that they are part of are also migrated successfully: for example, in the case of migrating an image embedded in a web page, not only must the image be migrated, but the integrity of the webpage must be maintained. This involves updating links to the image maintained by other information objects within the archive so that the migrated format is correctly linked from those objects.

How exactly these updates are managed depends on how the archive file is recompressed, as discussed below.

5.5 Validation of migration

Following migration each conceptual entity identified during conceptual characterization must be checked to ensure that it still exists, still has the same links to other conceptual entities and still has the same transformational information properties. For example, for images, these properties may include the histogram spread of red green and blue pixels while for documents it typically includes the number of pages. To validate successful migration, physical and conceptual characterization is performed on the migrated files. These properties are compared to the original, to confirm that they have not been changed by the migration, which would indicate a failure in migration.

5.6 Rewrapping content back in containers

At the end of the migration, it is important that a (W)ARC file is recreated so it can be utilized by the appropriate access workflows, e.g., in the Wayback machine [17]. This means that the (W)ARC containers need to be recreated using the appropriate combination of migrated and non-migrated files. As discussed earlier, to maintain the integrity of migrated web pages, links to the migrated files must also be updated appropriately.

The reconstruction of WARC files creates a specific practical problem: the specification for WARC files includes protocols for migrating files inside a WARC container and recording the details (provenance) of the migration in conversion record metadata in the WARC. However, most WARC access workflows, such as the Wayback machine, do not currently support conversion records, so WARC files migrated in this way will not be properly displayed. This required the development of two different workflows for creating migrated WARC files: one, which is formally correct according to the WARC standard, and maintains the integrity of the WARC schema, and a second which is more pragmatic, and produces a file that can be displayed correctly by current WARC viewers. This pragmatic workflow can also be used for the migration of container formats which do not support conversion records, such as ARC files.

If the formally correct workflow is chosen, then the workflow creates conversion records for each migrated file, which reference the original WARC file pre-migration. To reconstruct the full archive, both the original WARC file and the new WARC file containing the conversion records are required. Links and other references to converted files are not updated, as a strict implementation of the WARC viewer should be able to retrieve the most recent version of the updated files.

If the pragmatic workflow is chosen, then the archive simply replaces the unmigrated version of the file in the archive with the new, migrated version. To maintain the integrity of the migrated webpages, links are updated where possible to refer to the migrated files, for example, updating files extensions where necessary. A new archive file is created which contains migrated files, files modified because they reference migrated files and files from the original archive which have been unaffected by the whole process..

In either workflow, once the new (W)ARC containers have been created, they are re-ingested into the archive, and the associated archive object metadata is updated to reflect the provenance of the transformation action that has been performed.

5.7 Limitations on validation

The migration process involved the following conceptual steps;

1. Unpacking of the (W)ARC files
2. Migrating the at risk files and modifying affected files
3. Repacking of the (W)ARC files

Ideally it would be possible to directly compare the transformational information properties of the (W)ARC file as it exists before step 1 and the (W)ARC file existing after step 3. The original characterization does indeed take place before step 1 but it uses the same unpacking process as step 1 before characterizing so it is equivalent to taking place after step 1. The same is true in reverse for the second characterization step meaning that the only true verification of the above process is taking place by comparing properties produced before and after step 2. This is probably reasonable since the process of packing and unpacking (W)ARC files is unlikely to lead to information loss or data corruption. However, it might still be better to have alternative implementations of packing and unpacking in migration and in characterization so that the process could be independently verified.

6. CONCLUSION

By using the initial characterization process to aggregate information on the objects contained in a web archive, and by storing these aggregated properties as properties of the container object, we considerably reduce the number of entities that need to be searched to perform preservation actions, and hence increase the scalability of web preservation, while maintaining efficient characterization during ingest.

While described using file format migration as an illustrative example, this method is not limited to describing file formats: any property which can be aggregated across the archive could also be recorded and retrieved using this method.

While this approach was developed to deal with the challenges of large scale web crawls, it would also have advantages across a large number of other situations in which characterization and file format migration (or other, similar operations) need to be performed across embedded file formats, provided that suitable software tools are available for identification/validation of the container objects and extraction of the embedded files formats. For example, the same approach has been successfully applied to the preservation of other container formats, such as .zip files.

7. ACKNOWLEDGMENTS

This research was partly funded by the European Union as part of APARSEN - Alliance Permanent Access to the Records of Science in Europe Network- under FP7-ICT-2009 agreement 269977.

8. REFERENCES

[1] For a recent review see “Web Archiving”, Maureen Pennock, DPC Technology Watch Report, March 2013

[2] Clément Oury, Sébastien Peyrard. From the World Wide Web to Digital Library Stacks: Preserving the French Web Archives. In Proc. iPRES2011, Singapore, 2011.

[3] www.httrack.com/

[4] crawler.archive.org/index.html

[5] netarkivet.statsbiblioteket.dk/

[6] Robert Sharpe. Active Preservation of Web Sites. In Proc. International Web Archiving Workshop IWAW 2010, Vienna, 2010.

[7] http://archive.org/web/researcher/ArcFileFormat.php

[8] http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=44717

[9] Information from Sébastien Peyrard in panel session at iPres 2011.

[10] Strodl S., Beran P. and Rauber A. Migrating content in WARC files 2009 The 9th International Web Archiving Workshop (IWAW 2009) Proceedings", (2009), 43 - 49

[11] Rosenthal, David S.H.; (2010) "Format obsolescence: assessing the threat and the defenses, Library Hi Tech, Vol. 28 Iss: 2, pp.195 – 210

[12] Jackson, Formats over Time: Exploring UK Web History, arXiv:1210.1714

[13] http://aws.amazon.com/ec2/instance-types

[14] http://www.nationalarchives.gov.uk/information-management/our-services/dc-file-profiling-tool.htm /

[15] http://openplanetsfoundation.org/blogs/2013-01-09-year-fits

[16] Jason Pierson, Mark Evans, James Carr and Robert Sharpe. Considerations for High Throughput Digital Preservation. In Proc. iPRES2011, Singapore, 2011, Page 267

[17] http://archive.org/web/web.php

CLEAR: a credible method to evaluate website archivability

Vangelis Banos† Yunhyong Kim‡ Seamus Ross‡ Yannis Manolopoulos†

†Aristotle University of Thessaloniki, Greece
vbanos@gmail.com manolopo@csd.auth.gr

‡University of Glasgow, United Kingdom
{yunhyong.kim,seamus.ross}@glasgow.ac.uk

ABSTRACT

Web archiving is crucial to ensure that cultural, scientific and social heritage on the web remains accessible and usable over time. A key aspect of the web archiving process is optimal data extraction from target websites. This procedure is difficult for such reasons as, website complexity, plethora of underlying technologies and ultimately the open-ended nature of the web. The purpose of this work is to establish the notion of Website Archivability (WA) and to introduce the Credible Live Evaluation of Archive Readiness (CLEAR) method to measure WA for any website. Website Archivability captures the core aspects of a website crucial in diagnosing whether it has the potentiality to be archived with completeness and accuracy. An appreciation of the archivability of a web site should provide archivists with a valuable tool when assessing the possibilities of archiving material and influence web design professionals to consider the implications of their design decisions on the likelihood could be archived. A prototype application, archiveready.com, has been established to demonstrate the viabiity of the proposed method for assessing Website Archivability.

Categories and Subject Descriptors

H.3 [**Information Storage And Retrieval**]: Online Information Services—*Web-based services*; H.3.7 [**Digital Libraries**]: [Collection]

General Terms

Web Archiving, Website Evaluation Method

Keywords

Web Archiving, Digital Preservation, Website Archivability

1. INTRODUCTION

Web archiving is the process of gathering up digital materials from the World Wide Web, ingesting it, ensuring that these materials are preserved in an archive, and making the collected materials available for future use and research [16]. Web archiving is crucial to ensure that our digital materials remain accessible over time.

Web archiving has two key aspects: organizational and technical. The organizational aspect of web archiving involves the entity that is responsible for the process, its governance, funding, long term viability and personnel responsible for the web archiving tasks [21]. The technical aspect of web archiving involves the procedures of web content identification, acquisition, ingest, organization, access and use [5, 25].

In this work, we are addressing two of the main challenges associated with technical aspects of web archiving, the acquisition of web content and the quality assurance (QA) performed before it is ingested into a web archive. Web content acquisition and ingest is a critical step in the process of web archiving; if the initial Submission Information Package (SIP) lacks completeness and accuracy for any reason (e.g. missing or invalid web content), the rest of the preservation processes are rendered useless. In particular, QA is vital stage in ensuring that the acquired content is complete and accurate.

The peculiarity of web archiving systems in comparison to other archiving systems, is that the SIP is preceded by an automated extraction step. Websites often contain rich information not available on their surface. While the great variety and versatility of website structures, technologies and types of content is one of the strengths of the web, it is also a serious weakness.There is no guarantee that web bots dedicated to retrieving website content (perform web crawling) can access and retrieve website content successfully [9].

Websites benefit from following established best practices, international standards and web technologies if they are to be amenable to being archived. We define the sum of the attributes that make a website amenable to being archived as *Website Archivability*. This work aims to:

- Provide mechanisms to improve the quality of web archive content (e.g. facilitate access, enhance content integrity, identify core metadata gaps).
- Expand and optimize the knowledge and practices of web archivists, supporting them in their decision making, and risk management, processes.
- Standardize the web aggregation practices of web archives, especially in relation to QA.
- Foster good practices in website development and web content authoring that make sites more amenable to harvesting, ingesting, and preserving.
- Raise awareness among web professionals regarding web preservation.

In this work, we define the *Credible Live Evaluation of Archive Readiness (CLEAR) method*, a set of metrics to quantify the level of archivability of any website. This method is designed to consolidate, extend and complement empirical web aggregation practices through the formulation of a standard process to measure if a website is archivable. The main contributions of this work are:

- the introduction of the notion of Website Archivability,
- the definition of the Credible Live Evaluation of Archive Readiness (CLEAR) method to measure Website Archivability,
- the description of ArchiveReady.com, a web application which implements the proposed method.

The concept of CLEAR emerged from our current research in web preservation in the context of the BlogForever project[1] which involves weblog harvesting and archiving. Our work revealed the need for a method to assess website archive readiness in order to support web archiving workflows.

The remainder of this paper is organized as follows: Section 2 presents work related to web archiving, content aggregation and QA, Section 3 introduces and analyses the CLEAR method, Section 4 presents archiveready.com, a prototype web application implementing it, Section 5 discusses future work and, Section 6 summarises our results.

2. RELATED WORK AND CONTEXT

The web archiving workflow includes identification, appraisal and selection, acquisition, ingest, organization and storage, description and access [16]. This section focuses explicitly on the acquision of web content and the way it is handled by web archiving projects and initiatives.

Web content acquision is one of the most delicate aspects of the web archiving workflow because it depends heavily on external systems: the target websites, web servers, application servers, proxies and network infrastructure. The number of independent and dependent elements gives harvesting a substantial risk load.

Web content acquisition for web archiving is performed using robots, also known as “spiders”, “crawlers”, or “bots”, self-acting agents that navigate around-the-clock through the hyperlinks of the web, harvesting topical resources without human supervision [18]. The most popular web harvester, Heritrix is an open source, extensible, scalable, archival quality web crawler [15] developed by the Internet Archive[2] in partnership with a number of libraries and web archives from across the world. Heritrix is currently the main web harvesting application used by the International Internet Preservation Consortium (IIPC)[3] as well as numerous web archiving projects. Heritrix is being continuously developed and extended to improve its capacities for intelligent and adaptive crawling [7] or capture streaming media [10]. The Heritrix crawler was originally established for crawling general webpages that do not include substantial dynamic or complex content. In response other crawlers have been developed which aim to address some of Heritrix’s shortcomings. For instance, BlogForever [2] is utilizing blog specific technologies to preserve blogs. Also, the ArchivePress project is based explicitly on XML feeds produced by blog platforms to detect web content [20].

As websites become more sophisticated and complex, the difficulties that web bots face in harvesting them increase. For instance, some web bots have limited abilities to process GIS files, dynamic web content, or streaming media [16]. To overcome these obstacles, standards have been developed to make websites more amenable to harvesting by web bots. Two examples are the Sitemaps.xml and Robots.txt protocols. The Sitemap.xml[4] protocol, ’Simple Website Footprinting’, is a way to build a detailed picture of the structure and link architecture of a website [12]. Implementation of the Robots.txt protocol provide web bots with information about specific elements of a website and their access permissions [26]. Such protocols are not used universally.

Web content acquisition for archiving is only considered complete once the quality of the harvested material has been established. The entire web archiving workflow is often handled using special software, such as the open source software Web Curator Tool (WCT)[5], developed as a collaborative effort by the National Library of New Zealand and the British Library, at the instigation of the IIPC. WCT supports such web archiving processes as permissions, job scheduling, harvesting, quality review, and the collection of descriptive metadata. Focusing on quality review, when a harvest is complete, the harvest result is saved in the digital asset store, and the Target Instance is saved in the Harvested state [6]. The next step is for the Target Instance Owner to Quality Review the harvest. WCT operators perform this task manually. Moreover, according to the web archiving process followed by the National Library of New Zealand, after performing the harvests, the operators review and endorse or reject the harvested material; accepted material is then deposited in the repository [19]. A report from the Web-At-Risk project provides confirmation of this process. Operators must review the content thoroughly to determine if it can be harvested at all [8].

Recent efforts to deploy crowdsourced techniques to manage QA provides an indication of how significant the QA bottleneck is. The use of these approaches is not new, they were deployed by digitisation projects. The QA process followed by most web archives is time consuming and potentially complicated, depending on the volume of the site, the type of content hosted, and the technical structure. However, to quote the IIPC, ”it is conceivable that crowdsourcing could support targeted elements of the QA process. The comparative aspect of QA lends itself well to ‘quick wins’ for participants”[7].

[1] http://blogforever.eu
[2] http://archive.org
[3] http://netpreserve.org
[4] http://www.sitemaps.org/
[5] http://webcurator.sourceforge.net/
[6] http://webcurator.sourceforge.net/docs/1.5.2/Web\%20Curator\%20Tool\%20User\%20Manual\%20(WCT\%201.5.2).pdf
[7] http://www.netpreserve.org/sites/default/files/..

IIPC has also organized a Crowdsourcing Workshop in its 2012 General Assembly to explore how to involve users in developing and curating web archives. QA was indicated as one of the key tasks to be assigned to users: "The process of examining the characteristics of the websites captured by web crawling software, which is *largely manual in practice*, before making a decision as to whether a website has been successfully captured to become a valid archival copy" [8].

The previous literature shows that there is an agreement within the web archiving community that web content aggregation is challenging. QA is an essential stage in the web archiving workflow but currently the process requires human intervention and research into automating QA is in its infancy. The solution used by web archiving initiatives such as Archive-it[9] is to perform test crawls prior to archiving[10] but these suffer from, at least, two shortcomings: a) the test crawls require human intervention to evaluate the results, and b) they do not fully address such challenges as deep-level metadata usage and media file format validation.

Website archivability provides an approach to automating QA, by assessing the amenability of a website to being archived before any attempt is made to harvest it. This approach would provide considerable gains by reducing computational and network resource usage through not harvesting unharvestable sites and by saving on human QA of sites that could not be harvested above particular quality thresholds.

3. ARCHIVABILITY EVALUATION METHOD

The main aspects of the Credible Live Evaluation of Archive Readiness (CLEAR) method (Ver.1, as of 04/2013). After introducing the objectives of CLEAR and its key components, we provide further analysis of all its aspects.

3.1 Introduction to CLEAR

The CLEAR method proposes an approach to producing on-the-fly measurement of *Website archivability*. Website archivability is defined as the extent to which a website meets the conditions for the safe transfer of the its content to a web archive for preservation purposes. All web archives currently employ some form of crawler technology to collect the content of target websites. These all communicate through HTTP requests and responses, processes that are agnostic of the repository system of the archive. Information such as the unavailability of pages, and other errors, is accessible as part of this communication exchange, and could be used by the web archive to support archival decisions (e.g. regarding retention, risk management, and characterisation). Here we combine this kind of information with an evaluation of the website's compliance with recognised practices in digital curation (e.g. using adopted standards, validating formats, and assigning metadata) to generate a credible score representing the archivability of target websites. Website archivability must not be confused with website dependability, the former refers to the ability to archive a website while the latter is a system property that integrates such attributes as reliability, availability, safety, security, survivability and maintainability[1].

The main components of CLEAR are:

- Archivability Facets: the factors that come into play and need to be taken into account to calculate total website archivability (e.g. standards compliance).
- Website Attributes: the website elements analysed to assess the Archivability Facets (e.g. the HTML markup code).
- Evaluations: the tests executed on the website attributes (e.g. HTML code validation against the W3C HTML standards) and approach used to combine the test results to calculate the archivability metric.

Each of the CLEAR components will be examined with respect to aspects of web crawler technology (e.g. hyperlink validation; performance measure) and general digital curation practices (e.g. file format validation; use of metadata) to propose five core constituent facets of archivability (Section 3.2). We further describe the website attributes (e.g. HTML elements; hyperlinks) used to examine each archivability facet (Section 3.3), and, finally, propose a method for combining tests on these attributes (e.g. validation of image format) to produce a quantitative measure that represents the website's archivability (Section 3.4).

3.2 Archivability Facets

Website archivability can be measured from several different perspectives. Here, we have called these perspectives *Archivability Facets* (See Figure 1). The selection of these facets is motivated by a number of considerations. For example, whether there are verifiable guidelines to indicate what and where information is held at the target website and whether access is available and permitted (i.e. Accessibility, see Section 3.2.1); whether included information follows a common set of format and/or language specifications (i.e. Standards Compliance, see Section 3.2.2); the extent to which information is independent from external support (i.e. Cohesion, see Section 3.2.4); the level of extra information available about the content (i.e. Metadata Usage, see Section 3.2.5); and, whether server response time is below an acceptable threshold (i.e. Performance, see Section 3.2.3).

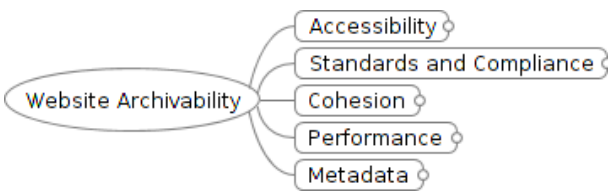


Figure 1: Archivability Facets: An Overview

3.2.1 F_A: Accessibility

A website is considered archivable only if web crawlers are able to visit its home page, traverse its content and retrieve

./CompleteCrowdsourcing.pdf

[8] http://netpreserve.org/sites/default/files/attachments/CrowdsourcingWebArchiving_WorkshopReport.pdf

[9] http://www.archive-it.org/

[10] https://webarchive.jira.com/wiki/display/ARIH/Test+Crawls

it via standard HTTP requests. In the case a crawler cannot find the location of all web resources, it will not be possible to retrieve the content. It is not only necessary to put resources on a web site, it is also essential to provide proper references to allow crawlers to discover them and retrieve them effectively and efficiently.

Example: a web developer is creating a website containing a javascript menu, which is generated on the fly. Web crawlers cannot understand this menu, so they are not able to find the web resources.

To support archivability, the website should, of course, provide valid links. In addition, a set of maps, guides, and updates for links should be provided to help crawlers find all the content (see Figure 2). These can be exposed in feeds, site maps, and robots.txt files. Information on whether the webpage is archived elsewhere (e.g. the Internet Archive[11]) and whether there are any errors in exporting them to the WARC format[12] could also help in determining the website accessibility.

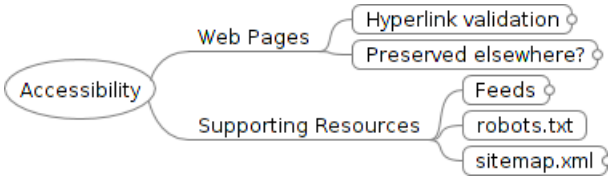


Figure 2: Archivability Facet: Accessibility

3.2.2 F_S: Standards Compliance

Compliance with standards is a recurring theme in digital curation practices (e.g. see Digital Preservation Coalition guidelines [4]). It is recommended that for digital resources to be preserved they need to be represented in known and transparent standards. The standards themselves could be proprietary, as long as they are widely adopted and well understood with supporting tools for validation and access. Above all, the standard should support disclosure, transparency, minimal external dependencies and no legal restrictions with respect to preservation processes that might take place within the archive[13].

Disclosure refers to the existence of complete documentation, so that, for example, file format validation processes can take place. Format validation is the process of determining whether a digital object meets the specifications for the format it purports to be. A key question in digital curation is, "I have an object purportedly of format F; is it really F?"[14]. Considerations of transparency and external dependencies refers to the resource's openness to basic tools (e.g. W3C HTML standard validation tool; JHOVE2 format validation tool).

[11] `http://www.archive.org`

[12] Popular standard archiving format for web content. `http://www.digitalpreservation.gov/formats/fdd/fdd000236.shtml`

[13] `http://www.digitalpreservation.gov/formats/sustain/sustain.shtml`

[14] `http://www.portico.org/digital-preservation/wp-content/uploads/2010/01/Portico_DLF_Fall2005.pdf`

Example: if a webpage has not been created using accepted standards, it is unlikely to be renderable by web browsers using established methods. Instead it is rendered in "Quirks mode", a custom technique to maintain compatibility with older/broken pages. The problem is that the quirks mode is really versatile. As a result, you cannot depend on it to have a standard rendering of the web site in the future.

We recommend validation be performed for three types of content (see Figure 3): webpage components (e.g. HTML and CSS), reference media content (e.g. audio, video, image, documents), and supporting resources (e.g. robots.txt, sitemap.xml, javascript).

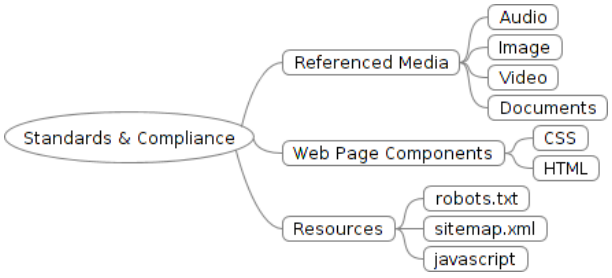


Figure 3: Archivability Facet: Compliance Standards

3.2.3 F_P: Performance

Performance is an important aspect of web archiving. The throughput of data acquisition of a web spider directly affects the number and complexity of web resources it is able to process. The faster the performance, the faster the ingestion of web content, improving a website's archiving process.

Example: if the performance of a website is slow, web spiders will have difficulty aggregating content and they may even abort if the performance degrates below a specific threshold.

While crawler performance can be adjusted and improved from within the archive, the server response time is under the control of the website creators. Website archivability is improved by optimising this response time. Depending on the size of the web archive and demands on acceptable server response time will differ. The performance is measured in relation to these needs. In a real world scenario, each archive would have a threshold indicating the maximum allowable server response time.

3.2.4 F_C: Cohesion

Cohesion is relevant for both the efficient operation of web crawlers, and, also, the management of dependencies within digital curation (e.g. see NDIIPP comment on format dependencies [17]). If files constituting a single website are dispersed across different services (e.g. different servers for images, javascript widgets, other resources), the acquisition and ingest is likely to risk suffering from neither being complete nor accurate. If one of the multiple services fails, the website fails. Here we characterise the robustness of the website in comparison to this kind of failure as *C*ohesion.

Example: images used in a website but hosted elsewhere

may cause problems in web archiving because they may not be captured when the site is archived. What is more, if the target site depends on 3rd party sites, the future availability of which is unknown, new kinds of problems are likely to arise.

The premise is that, keeping information associated to the same website together (e.g. using the same host for a single instantiation of the website content) would lead to a robustness of resources preserved against changes that occur outside of the website (cf. *encapsulation*[15]). Cohesion is tested on three levels:

- examining how many hosts are employed in relation to the location of referenced media content,
- examining how many hosts are employed in relation to supporting resources (e.g. robots.txt, sitemap.xml, and javascripts),
- examining the number of times proprietary software or plugins are referenced.

3.2.5 F_M: Metadata Usage

The adequate provision of metadata (e.g. see Digital Curation Centre Curation Reference Manual chapters on metadata [14], preservation metadata [23], archival metadata [27], and learning object metadata [11]) has been a continuing concern within digital curation (e.g. see seminal article by Lavoie[16] and insightful discussions going beyond preservation[17]). The lack of metadata impairs the archive's ability to manage, organise, retrieve and interact with content effectively. It is, widely recognised that it makes understanding the context of the material a challenge.

We will consider metadata on three levels (summarised in Figure 4). To avoid the dangers associated with committing to any specific metadata model, we have adopted a general view point shared across many information disciplines (e.g. philosophy, linguistics, computer sciences) based on syntax (e.g. how is it expressed), semantics (e.g. what is it about) and pragmatics (e.g. what can you do with it). There are extensive discussions on metadata classification depending on their application (e.g. see NISO classification [22]; discussion in DCC Curation Reference Manual chapter on Metadata [14]). Here we avoid these fine-grained discussions and focus on the fact that much of the metadata approaches examined in existing literature can be exposed already at the time that websites are created and disseminated.

For example, metadata such as transfer and content encoding can be included by the server in HTTP headers. The required end-user language to understand the content can be indicated as part of the HTML element attribute. Descriptive information (e.g. author, keywords) that can help understand how the content is classified can be included in the HTML META element attribute and values. Metadata

[15] http://www.paradigm.ac.uk/workbook/preservation-strategies/selecting-other.html

[16] http://www.dlib.org/dlib/april04/lavoie/04lavoie.html

[17] http://www.activearchive.com/content/what-about-metadata

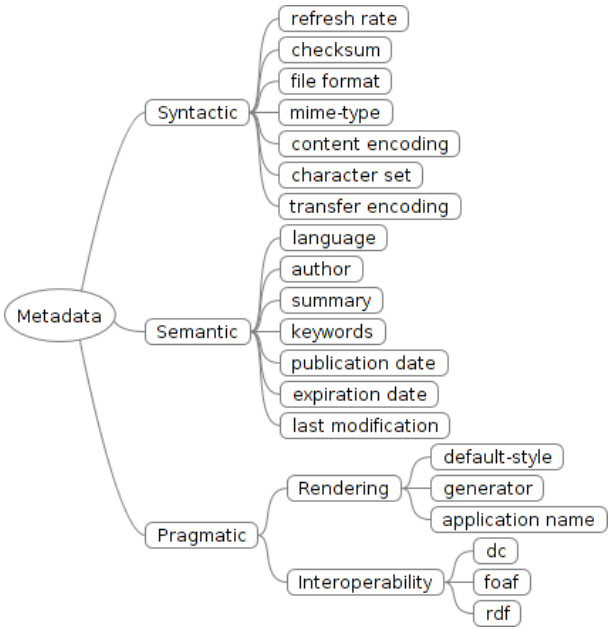


Figure 4: Archivability Facet: Metadata

that support rendering information, such as application and generator names, can also be included in the HTML META element. The use of other well known metadata and description schemas (e.g. Dublin Core [28]; Friend of a Friend (FOAF) [3]; Resource Description Framework (RDF) [13]) can be included to promote better interoperability. The existence of selected metadata elements can be checked as a way of increasing the probability of implementing automated extraction and refinement of metadata at harvest, ingest, or subsequent stage of repository management.

3.3 Websites Attributes

In this section, we examine the website attributes used to measure the archivability facets discussed in Section 3.2. In Figure 5, we have illustrated the components of the website that will be examined to measure the website's potential for meeting the requirements of the archivability facets.

For example, the level of **Accessibility** can be quantified on the basis of: whether or not,

- feeds exist (e.g. RSS and ATOM);
- robots.txt exists;
- sitemap.xml is mentioned in robots.txt and sitemap.xml exists at the location specified, and/or sitemap.xml is found at the root directory of the server;
- hyperlinks are valid and accessible; and,
- there are existing instantiations of the webpage elsewhere (e.g. snapshots at the Internet Archive[18]).

[18] http://www.archive.org

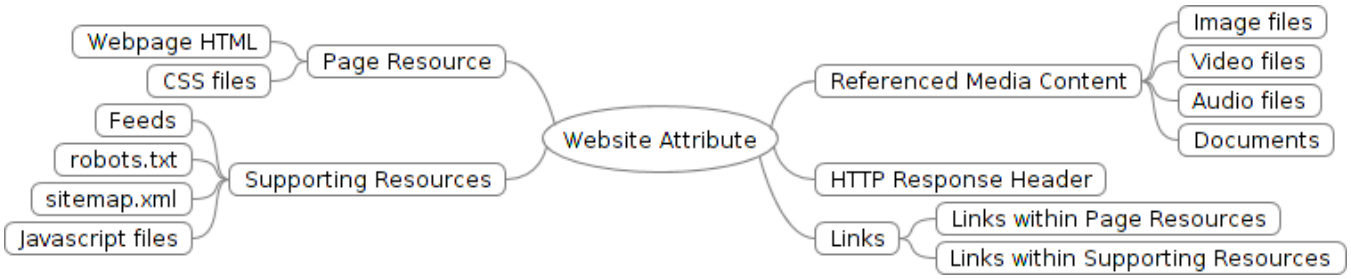


Figure 5: Website Archivability: mapping archivability facets to website attributes.

The existence of an RSS feed allows the publication of webpage content that can be automatically syndicated or exposed. It allows web crawlers automatically to retrieve updated content and the standardised format of the feeds allow access by many different applications. For example the BBC uses feeds to let readers see when new content has been added[19].

The file robots.txt[20] indicates to a web crawler which URLs it is allowed to crawl. The use of robots.txt helps preventing the retrieval of website content that would be aligned with permissions and special rights associated to the webpage.

The Sitemaps protocol, supported jointly by the most widely used search engines to help content creators and search engines, is an increasingly widely used way unlock this hidden data by making it available to search engines [24]. To implement the Sitemaps protocol, the file sitemap.xml is used to list all the pages of the website and their location. The location of this sitemap, if it exists, can be indicated in the robots.txt. Regardless of its inclusion in the robots.txt file, the sitemap, if it exists, should, ideally, be called 'sitemap.xml' and put at the root of your web server (e.g. `http://www.example.co.uk/sitemap.xml`).

The hyperlinks of the website can be examined for availability as an indication of website accessibility. A website with many missing and/or broken links is not likely to be archived to any significant degree of completeness or accuracy.

The website will be checked for **Standards Compliance** on three levels: referenced media format (e.g. image and audio included in the webpage), webpage (e.g. HTML and CSS markup) and resource (e.g. sitemap, scripts). Each one of these are expressed using a set of specified file formats and/or languages. The languages (e.g. XML, javascript) and formats (e.g. jpeg) will be validated using tools, such as W3C HTML[21] and CSS validator[22], JHOVE2[23] and/or Apache Tika[24] file format validator, python XML validator[25], robots.txt checker[26], ECMAScript[27] language specification.

The level of **Cohesion** is measured by the extent to which material associated to the website is kept within one host. This is measured by the proportion of content, resources, and plugins that are sourced internally. This can be examined through an analysis of links, on the level of referenced media content, and on the level of supporting resources (e.g. javacsript). In addition the proportion of content relying on predefined proprietary software can be assessed and monitored.

The calculation of **Performance** is straightforward based on the response time of the server and can be implemented as a pass/fail test depending on a pre-set threshold of acceptability. In a archival context, it is likely that there is an acceptable performance threshold for the website if it is to be archivable given the web crawler and archival objectives.

The score for **Metadata Usage** can be assessed on the basis of whether or not,

- the <HTML> element includes a "lang" attribute specifying a value for the primary end-user language;
- the website includes element tags (i.e. <dc>, <foaf>, <rdf>), that indicate the use of Dublin Core, FOAF, and RDF (in the long-term, other elements related to initiatives such as SIOC[28], LOD[29], ORE[30] can be added as needed);
- fixity information is included ("content-md5" attribute can be used to include this in the HTTP response header);
- content mime-type identification is available ("content-type" can be used in the HTTP response header to indicate this; in cases where it is missing, this process might be refined to use JHOVE2 or Apache Tika to identify the format of content);
- character set is described (this can be exposed using "content-type" along with mime-type);
- transfer encoding is specified (this describes and compression methods in use and can be specified in the HTTP response header);

[19] `http://www.bbc.co.uk/news/10628494`
[20] `http://www.robotstxt.org/`
[21] `http://validator.w3.org/`
[22] `http://jigsaw.w3.org/css-validator/`
[23] `http://www.jhove2.org`
[24] `http://tika.apache.org/`
[25] `http://code.google.com/p/pyxmlcheck/`
[26] `http://tool.motoricerca.info/robots-checker.phtml`
[27] `http://www.ecmascript.org/`
[28] `http://sioc-project.org/`
[29] `http://linkeddata.org/`
[30] `http://www.openarchives.org/ore/1.0/datamodel`

- content encoding is specified (this can be included in the HTTP response header);
- HTML <META> element includes: "author", "description", "keywords", "default-style", "application-name", "generator" & "refresh" information.

In the case of HTTP response header, the availability of selected metadata elements and their values will be examined. In the case of more specific metadata schemas such as DC, at this stage, we envision only examining whether the schema is being used or not. At a later stage we might extend this to examine which elements are in use. The premise is that the information in the HTTP response header is essential to the archivability, whereas the elements and values associated with specific standards are considered to be desirable characteristics that would lead to richer metadata generation but are not necessarily essential.

The <META> tag attribute often embodies semantic data (e.g. authorship and keywords); however, the quality of metadata here can vary widely. Metadata harvested from this element should be used in conjunction with that derived from other components, for example, RSS feeds and Microformats[31], where these are available.

3.4 Evaluations

Combining the information discussed in Section 3.3 to calculate a score for website archivability goes through the following steps.

- The website's archivability potential with respect to each facet will be represented by an N-tuple (x_1, ..., x_k, ..., x_N) where the value of x_k is a zero or one representing a negative or positive answer, respectively, to the binary question asked about that facet, and where N is the total number of questions associated to that facet. For example, an example question in the case of the Standards Compliance Facet would be "I have an object purportedly of format F; is it?"[32]; if there are M files for which format validation is being carried out then there will be M binary questions of this type.
- If all questions are considered to be of equal value to the facet, then the archivability with respect to the facet in questions is just the sum of all the coordinates divided by N (simplest model). If some questions are considered to be more important, then these can be assigned higher weights so that the archivability is $\sum_{k=0}^{N} \frac{\omega_k x_k}{N}$, where ω_k is the weight assigned to question k and $\sum \omega_k = 1$.
- If selected questions are grouped to represent sub-facets to be calculated at different hierarchical levels then this will also change the weighting. Ideally, this could be adjusted on the basis of the needs of the community for which the website is being archived. Some will be more interested in preservation of images, while others will be interested in text. This can be easily incorporated into the current methodology.

Once the archivability with respect to each facet is calculated, the total measure of Website Archivability can be simply defined as:

$$\sum_{\lambda \in \{A,S,C,P,M\}} w_\lambda F_\lambda$$

where F_A, F_S, F_C, F_P, F_M are archivability with respect to Accessibility, Standards Compliance, Cohesion, Performance, Metadata Usage, respectively, and $\sum_{\lambda \in \{A,S,C,P,M\}} w_\lambda = 1$ and $0 \leq w_\lambda \leq 1 (\lambda \in \{A, S, C, P, M\})$.

Depending on the curation and preservation objectives of the web archive, the weight of each facet is likely to be different, and w_λ should be assigned to reflect this. In the simplest model, these can be set to be equal so that $w_\lambda = 0.2$ for all λ. In actuality accessibility will be the most central consideration in archivability since, if the content cannot be found or accessed, then the website's compliance with other standards, and conditions become moot.

4. A WEBSITE ARCHIVABILITY EVALUATION TOOL: ARCHIVEREADY.COM

ArchiveReady, a web application located at `http://www.archiveready.com`, implements the CLEAR method for evaluating website archivability. We describe its technology stack, and website archivability evaluation workflow. To demonstrate ArchiveReady, we also present an evaluation of the iPRES2013 Conference website.

4.1 Technology Stack

ArchiveReady is a web application based on the following key components: Debian linux[33] operating system for development and production servers, Nginx web server[34] to server static web content, Python programming language[35], Gunicorn python WSGI HTTP Server for unix[36] to server dynamic content, BeautifulSoup[37] to analyse html markup and locate elements, Flask[38], a python microframework to develop web applications, Redis advanced key-value store[39] to manage job queues and temporary data, Percona Mysql RDBMS[40] to store long-term data. JSTOR/Harvard Object Validation Environment (JHOVE) [6] for object validation, Javascript libraries such as jQuery[41] and Bootstrap[42] are utilized to create a compeling user interface.

To ensure high level compatibility with W3C standards the initiative used open source web services provided by the

[31] `http://microformats.org/about`
[32] `http://www.portico.org/digital-preservation/wp-content/uploads/2010/01/Portico_DLF_Fall2005.pdf`
[33] `http://www.debian.org`
[34] `http://www.nginx.org`
[35] `http://www.python.org/`
[36] `http://gunicorn.org/`
[37] `http://www.crummy.com/software/BeautifulSoup/`
[38] `http://flask.pocoo.org/`
[39] `http://redis.io`
[40] `http://www.percona.com`
[41] `http://www.jquery.com`
[42] `http://twitter.github.com/bootstrap/`

W3C. These include: the Markup Validator[43], the Feed Validation Service[44] and the CSS Validation Service[45].

The greatest challenge in implementing ArchiveReady is performance. According to the HTTP Archive Trends, the average number of HTTP requests initiated when accessing a web page is over 90 and is expected to rise[46]. In response to this performance context, ArchiveReady has to be capable of performing a very large number of HTTP requests, process the data and present the outcomes to the user in real time. This is not possible with a single process for each user, the typical approach in web applications. To resolve this blocking issue, an asynchronous job queue system based on Redis for queue management and the Python RQ library[47] was deployed. This approach enables the parallel execution of multiple evaluation processes, resulting in huge performance benefits when compaired to traditional web application execution model.

4.2 Workflow

ArchiveReady is a web application providing two types of interaction: web interface and web service. With the exception of presentation of outcomes (HTML for the former and JSON for the latter) both are identical. The workflow can be summarised as follows:

1. ArchiveReady receives a target URL and performs an HTTP request to retrieve the webpage hypertext.

2. After analysing it, multiple HTTP connections are initiated in parallel to retrieve all web resources referenced in the target webpage, imitating a web spider. ArchiveReady analyses only the URL submitted by the user, it does not evaluate the whole website recursively.

3. In stage 3, Website Attributes are evaluated (See Section 3.3).

4. The metrics for the Archivability Facets are calculated according to the CLEAR method and the final website archivability rating is calculated.

Note that in the current implementation, CLEAR evaluates only a single webpage based on the assumption that all website pages share the same components, standards and technologies. This issue is further discussed in Future Work.

4.3 Demonstration

To demonstrate ArchiveReady, we evaluate the website of iPRES'2013 international conference as it was available on 23 April 2013[48] and present the results in Table 1. The corresponding result is also presented in Figure 6.

[43] `http://validator.w3.org/`
[44] `http://validator.w3.org/feed/`
[45] `http://jigsaw.w3.org/css-validator/`
[46] `http://httparchive.org/trends.php`
[47] `http://python-rq.org/`
[48] `http://ipres2013.ist.utl.pt/`

Table 1: http://ipres2013.ist.utl.pt/ Website Archivability evaluation

Facet	Evaluation	Rating	Total
Accessibility	No RSS feed	50%	50%
	No robots.txt	50%	
	No sitemaps.xml	0	
	6 Valid links	100%	
Cohesion	1 external & no internal scripts	0	70%
	4 local & 1 external images	80%	
	No QuickTime or Flash objects	100%	
	1 local CSS file	100%	
Standards Compliance	1 Invalid CSS file	0	77%
	Invalid HTML	0	
	Meta description found	100%	
	No content encoding in HTTP headers	50%	
	Content type HTTP header	100%	
	Page expiration HTTP header	100%	
	Last-modified HTTP header	100%	
	No QuickTime or Flash objects	100%	
	5 images checked successfully with JHOVE	100%	
Metadata	Meta description found	100%	87%
	Content type	100%	
	No page expiration metadata	50%	
	Last-modified HTTP header	100%	
Performance	Avg network response time is 0.546ms	100%	100%
		Website Archivability	77%

5. FUTURE WORK

Future work directions stem from two facts: a) the identification of limitations which nevertheless do not refute the claim that the proposed method is significant, and b) the novelty of this work which promises to improve considerably the web archiving process.

The method as currently implemented treats all website archivability facets equally in calculating the total Archivability Score. This may not be the optimal approach as in different organisational and policy contexts the objectives of web archiving might put greater or lesser emphasis on the individual Archivability Facets. The ability to weight the various individual Archivability Facet Scores in calculating the total Archivability Score is a feature which users will find valuable. For instance Metadata breadth and depth might be critical for a particular web archiving research task and

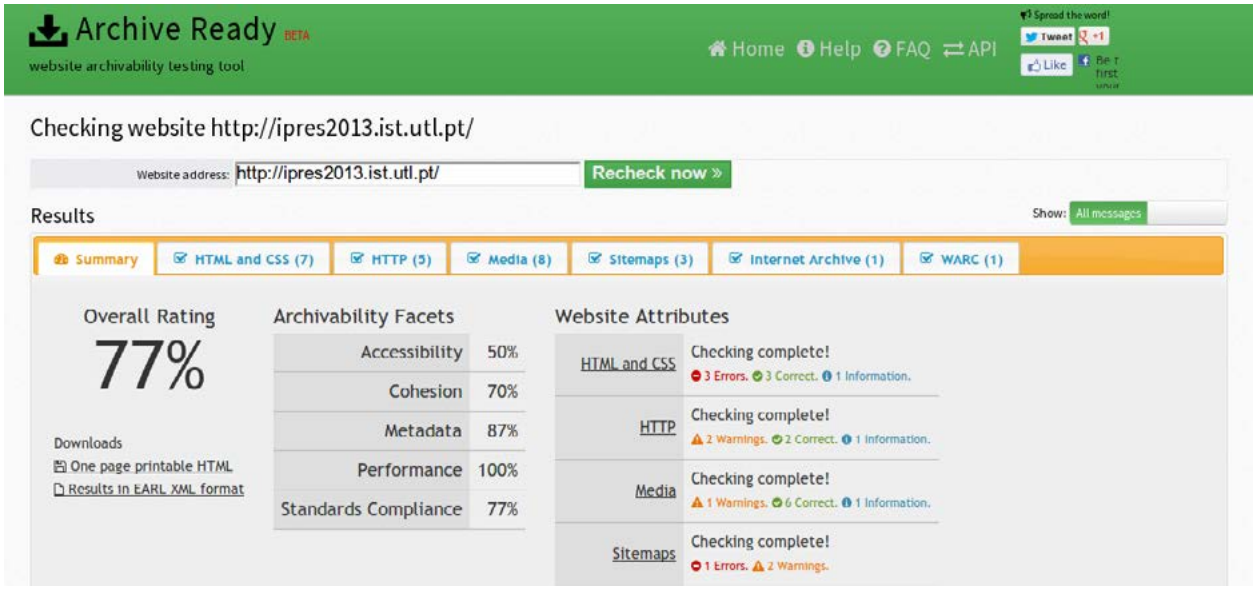


Figure 6: Evaluating iPRES2013 Website Archivability using ArchiveReady

therefore in establishing the archivability score for a particular site the user may which to instantiate this thinking in calculating the overall score. A next step will be to introduce a mechanism to allow the user to weight each Archivability Facet to reflect specific objectives.

Currently, CLEAR evaluates only a single website page based on the assumption that webpages from the same website share the same components and standards. To achieve a more objective evaluation, it would be better to perform sampling using sitemap.xml and RSS referenced pages to increase the breadth of the target website content to be evaluated.

There are some open questions that could lead to further refinement of the website archivability concept:

- Is it correct to consider archivability to be directly proportional to the number of binary questions answered positively? Are there points in the archivability curve that move at a faster/slower rate?
- Evidence from other archiving projects demonstrates that certain classes and specific types of errors create lesser or greater obstacles to website acquisition and ingest than others. The website archivabliity tool needs to be enhanced to reflect this differential valuing of error classes and types.
- Recognising that the different classes and types of errors do not have a purely summative combinatorial impact on archivability of a website this research in its next stage must identify the optimal way to reflect this weighting to enable comparisons across websites.
- Currently the system is envisaged as being used to guide the process of archiving websites, but a further extension would support its use by developers to assist them in design and implementation.

One way to address these concerns might be to apply an approach similar to normalized discounted cummulative gain (NDCG) in information retrieval[49]: for example, a user can rank the questions/errors to prioritise them for each facet. The basic archivability score can be adjusted to penalise the outcome when the website does not meet the higher ranked criteria. Further experimentation with the tool will lead to a richer understanding of new directions in automation in web archiving.

6. CONCLUSIONS

Our main aims were to improve web archive quality by establishing standards and tools to enhance content aggregation. Moreover, our aims were to help web archive operators improve their content ingestion workflows and also raise awareness among web professionals regarding web archiving.

To this end, we introduced the *Credible Live Evaluation of Archive Readiness (CLEAR)* method, a set of metrics to quantify the level of *Website Archivability* based on established web archiving standards, digital preservation principles and good practices. Also, one of the authors of this paper developed a web application implementing this method, ArchiveReady.com. This approach, provided the authors with an environment to test the concept of Archivability Facets and offered a method for web archive operators to evaluate target websites before content harvesting and ingestion, thus avoiding invalid harvests, erroneous web archives and unnecessary wasted resources which could be used elsewhere. ArchiveReady provides web professionals with an easy but thorough tool to evaluate their websites and improve their archivability. This achieved the twin goals of on the one hand instantiating methods to improve website archiving and on the other raising awareness of the challenges to web archiving among a broader audience.

[49] `http://en.wikipedia.org/wiki/Discounted_cumulative_gain\#Normalized_DCG`

7. ACKNOWLEDGMENTS

The research leading to these results has received funding from the European Commission Framework Programme 7 (FP7), BlogForever project, grant agreement No.269963.

8. REFERENCES

[1] A. Avižienis, J.-C. Laprie, and B. Randell. Fundamental concepts of computer system dependability. In *Proceedings of IARP/IEEE-RAS Workshop on Robot Dependability: Technological Challenge of Dependable, Robots in Human Environments*, 2001.

[2] V. Banos, N. Baltas, and Y. Manolopoulos. Trends in blog preservation. In *Proceedings of the 14th International Conference on Enterprise Information Systems (ICEIS)*, Wroclaw, Poland, 2012.

[3] D. Brickley and L. Miller. Foaf vocabulary specification 0.98. *Namespace Document*, 9, 2010.

[4] D. P. Coalition. Institutional strategies - standards and best practice guidelines. `http://www.dpconline.org/advice/preservationhandbook/institutional-strategies/standards-and-best-practice-guidelines`, 2012. [Online; accessed 18-April-2013].

[5] D. Denev, A. Mazeika, M. Spaniol, and G. Weikum. The sharc framework for data quality in web archiving. *The VLDB Journal*, 20(2):183–207, 2011.

[6] M. Donnelly. Jstor/harvard object validation environment (jhove). *Digital Curation Centre Case Studies and Interviews*, 2006.

[7] M. Faheem and P. Senellart. Intelligent and adaptive crawling of web applications for web archiving. In *Proceedings of the 21st International Conference Companion on World Wide Web (WWW)*, pages 127–132, Lyon, France, 2012.

[8] V. D. Glenn. Preserving government and political information: The web-at-risk project. *First Monday*, 12(7), 2007.

[9] Y. He, D. Xin, V. Ganti, S. Rajaraman, and N. Shah. Crawling deep web entity pages. In *Proceedings of the 6th ACM International Conference on Web Search and Data Mining (WSDM)*, pages 355–364, Rome, Italy, 2013.

[10] H. Hockx-Yu, L. Crawford, R. Coram, and S. Johnson. Capturing and replaying streaming media in a web archive-a british library case study, 2010.

[11] U. o. S. Lorna Campbell. Learning object metadata, curation reference manual. `http://www.dcc.ac.uk/resources/curation-reference-manual/completed-chapters/learning-object-metadata`, 2007. [Online; accessed 18-April-2013].

[12] S. Mansfield-Devine. Simple website footprinting. *Network Security*, 2009(4):7–9, 2009.

[13] B. McBride et al. The resource description framework (rdf) and its vocabulary description language rdfs. *Handbook on Ontologies*, pages 51–66, 2004.

[14] D. Michael Day. Metadata, curation reference manual. `http://www.dcc.ac.uk/resources/curation-reference-manual/completed-chapters/metadata`, 2005. [Online; accessed 18-April-2013].

[15] G. Mohr, M. Stack, I. Rnitovic, D. Avery, and M. Kimpton. Introduction to heritrix. In *Proceedings of the 4th International Web Archiving Workshop (IWAW)*, Vienna, Austria, 2004.

[16] J. Niu. An overview of web archiving. *D-Lib Magazine*, 18(3):2, 2012.

[17] L. of Congress. Sustainability of digital formats planning for library of congress collections: External dependencies. `http://www.digitalpreservation.gov/formats/sustain/sustain.shtml\#external`, 2013. [Online; accessed 18-April-2013].

[18] G. Pant, P. Srinivasan, and F. Menczer. Crawling the web. In *Web Dynamics*, pages 153–177. Springer, 2004.

[19] G. Paynter, S. Joe, V. Lala, and G. Lee. A year of selective web archiving with the web curator tool at the national library of new zealand. *D-Lib Magazine*, 14(5):2, 2008.

[20] M. Pennock and R. Davis. Archivepress: A really simple solution to archiving blog content. In *Proceedings of the 6th International Conference on Preservation of Digital Objects (IPres)*, San Francisco, CA, 2009.

[21] M. Pennock and B. Kelly. Archiving web site resources: a records management view. In *Proceedings of the 15th International Conference on World Wide Web (WWW)*, pages 987–988, Edinburgh, UK, 2006.

[22] N. Press. Understanding metadata. *National Information Standards*, 20, 2004.

[23] F. C. f. L. A. Priscilla Caplan, Digital Library Services. Preservation metadata, curation reference manual. `http://www.dcc.ac.uk/resources/curation-reference-manual/completed-chapters/preservation-metadata`, 2006. [Online; accessed 18-April-2013].

[24] U. Schonfeld and N. Shivakumar. Sitemaps: above and beyond the crawl of duty. In *Proceedings of the 18th International Conference on World Wide Web (WWW)*, pages 991–1000, Madrid, Spain, 2009.

[25] M. Spaniol, D. Denev, A. Mazeika, G. Weikum, and P. Senellart. Data quality in web archiving. In *Proceedings of the 3rd Workshop on Information Credibility on the Web (WICOW)*, pages 19–26, Madrid, Spain, 2009.

[26] Y. Sun, Z. Zhuang, and C. L. Giles. A large-scale study of robots. txt. In *Proceedings of the 16th International Conference on World Wide Web (WWW)*, pages 1123–1124, Banf, Canada, 2007.

[27] W. D. . M. van Ballegooie. Archival metadata, curation reference manual. `http://www.dcc.ac.uk/resources/curation-reference-manual/completed-chapters/archival-metadata`, 2006. [Online; accessed 18-April-2013].

[28] S. Weibel, J. Kunze, C. Lagoze, and M. Wolf. Dublin core metadata for resource discovery. *Internet Engineering Task Force RFC*, 2413:222, 1998.

Interoperability of web archives and digital libraries: A Delphi study

Hendrik Kalb
Institute for Business Informatics
Technische Universität Berlin
Berlin, Germany
hendrik.kalb@tu-berlin.de

Paraskevi Lazaridou
Institute for Business Informatics
Technische Universität Berlin
Berlin, Germany
paraskevi.lazaridou@tu-berlin.de

Edward Pinsent
Academic & Research Technologies
University of London Computer Centre
London, UK
edward.pinsent@london.ac.uk

Matthias Trier
Department of IT Management
Copenhagen Business School
Copenhagen, Denmark
mt.itm@cbs.dk

ABSTRACT

The interoperability of web archives and digital libraries is crucial to avoid silos of preserved data and content. While various researches focus on specific facets of the challenge to interoperate, there is a lack of empirical work about the overall situation of actual challenges. We conduct a Delphi study to survey and reveal the insights of experts in the field. Results of our study are presented in this paper to enhance further research and development efforts for interoperability.

Categories and Subject Descriptors

H.3.7 [**Digital Libraries**]; D.2.12 [**Interoperability**]

General Terms

Study, Theory

Keywords

Interoperability, Web Archives, Digital Libraries, Delphi Study

1. INTRODUCTION

This paper reveals the results of a Delphi study about the interoperability of web archives and digital libraries. The aim of the study is to provide a better understanding about crucial aspects of interoperability in this domain.

According to the Institute of Electrical and Electronics Engineers (IEEE), interoperability is the "ability of two or more systems or components to exchange information and to use the information that has been exchanged" [7, p. 114]. It has numerous facets including uniform naming, metadata formats, document models, and access protocols [16]. Interoperability in a narrow sense describes how technical systems interoperate. In a broader sense, it also comprises social, political, and organisational factors [9].

Research about interoperability of web archives and digital libraries often considers technical and semantic aspects of interoperability, e.g. protocols and standards. However, conceptual models for interoperability also comprise other aspects. The European Interoperability Framework differentiates between organisational, semantic, and technical interoperability [12]. This perspective has been adopted by the DL.org to describe and consider interoperability of digital libraries [4]. Other publications mention the semantic level under the term content level [3]. A similar perspective distinguishes between technical / basic, syntactic, functional / pragmatic, and semantic as levels with increasing abstraction [10]. Similarly, [25] describes with an increasing capability for interoperation the levels of technical, syntactic, semantic, pragmatic, dynamic, and conceptual interoperability. A specific perspective for digital libraries distinguishes the levels of gathering, harvesting, and federation [3]. The maturity of inter-organisational interoperability can be assessed on the levels of computer, process, knowledge, value, and goal interoperability [9]. While the different models indicate interoperability as a subject with various facets, only a few studies can be found that inquire into current web archives and digital libraries for interoperability issues (see Section 2). Thus, there is a risk that research and development to improve interoperability is mainly driven by personal assumptions, beliefs, or experiences of the researcher, and therefore fails to address the real needs of the community.

Our survey aims to gain insight into areas that have not been surveyed and derives from people who are highly involved and have personal experience. Our aim is to examine a theoretical framework of interoperability in both web archives and digital libraries with the assistance of people who have their own experiences and thoughts on the

topic. The survey does not focus on specific technical developments and offers to the participant the opportunity to develop freely their thoughts. This study can be considered as a discussion about interoperability; the obstacles, the current limitations, the followed approaches, the forthcoming challenges, the ideas for improvement. Therefore, our contribution, not only to the research community but as well to the involved communities, is the sharing of the valuable outcome of an enlightening virtual discussion from experts about interoperability.

The remainder of this paper is organised as follows: Section 2 reviews related studies about web archives and digital libraries, and focuses on their relation to interoperability aspects. In section 3, we reveal the chosen method for our study including a short introduction into the Delphi method in general. The first and second round's results of our study are presented in section 4 and 5 before we discuss them in section 6. Finally, we draw our conclusions in section 7.

2. RELATED WORK

In this section, we review published surveys on web archives and digital libraries regarding their insights into interoperability aspects.

Several surveys have been conducted in the domains of web archiving and digital libraries addressing issues regarding the web archiving activities. The majority put their focus on web archiving in general, examining the progress that has been made, the maturity so far, the problems encountered and the practices used in the field of web archiving. However, interoperability was out of their scope. Aspects such as legislation and national legal authorities, permission access, common tools and standards and cooperation of institutions for common developments that are also addressed, are of course related to our topic of research but not in a direct way. In particular, the International Internet Preservation Consortium (IIPC) carried out a survey among its members, basically a profile identification, and got responses from 35 of them. The survey examined the maturity of web archiving, the scope, the tools used for harvesting, curation and access, legal limitations by their countries and access restrictions [11]. Another survey on 16 national libraries focuses on how they attempt to justify their web archiving activities [23]. The Internet Memory Foundation ran a survey on European institutions aiming to obtain a clearer understanding of problems encountered in the field of Internet archiving with the help of a panel composed of 74 participants from national libraries, audiovisual and institutional archives. This survey addressed several aspects such as the status of web archiving, legal aspects, access restrictions, policies and priorities regarding the scope and the types of archiving [1]. The 18th Conference of Directors of National Libraries in Asia and Oceania (CDNLAO) presented a report with the participants' answers about web archiving in this region. The questions were about cooperation, access and preservation policies, tools in use, and the legal framework [20]. Later, a survey presented an updated overview of the web archiving initiatives internationally, in which the addressed aspects were mainly the scope, content characteristics, file formats, technologies and the provided access [8].

However, to the best of our knowledge, only a few surveys put their focus specifically on interoperability issues. A study regarding the future interoperability in web archiving was presented by [14] as a survey on national libraries. In this study, 37 participants responded to questions regarding several issues like scope of harvested resources, collecting and discovering policies, level of harvesting, access to archived content, level of cooperation with other web archives, how they solve ownership and technical issues and what kind of institutions they could partner with to solve such problems. The motivation of this survey, and also one of the questions, was the belief that interoperability between all national web archives should be a long-range goal, and the majority agreed on that. Based on the results, it is clear that a great challenge and need for the national libraries is to make legal deposit, copyright and related legislation adapted to the world of the Internet so that the digital part of national heritage can be preserved for future generations. Additionally, some comments of participants regarding preferable engagement in partnerships revealed a need for cooperation with institutions that have to offer technical and collection expertise along with a commitment to preservation issues [14].

Another survey, that focused exclusively on interoperability aspects, but specifically in the digital library sphere, was conducted by the DL.org Policy Working Group [13]. This experimental survey on policy interoperability of digital libraries was carried out among a selected sample of digital libraries, digital repositories and federated services, and received 26 responses (15 completed). This survey addressed how the policies, strategies, frameworks and plans of the digital libraries affect or are affected by interoperability. Their findings revealed that existing policies of the organisations have been revised according to those of other organisations with regard to policy exchange and reuse only in the areas of Preservation, Access, Collection Development and Metadata. Furthermore, even if respondents expressed interest to interoperate with other public or private organisations, just few of the stated policies regulate such interactions. The authors identified in the results an indication for approaching policy interoperability not only from a technical but also from an organisational and semantic perspective [13]. Within the same project, another survey [26] was run regarding quality interoperability in digital libraries, since quality and interoperability are two aspects that affect each other. The results revealed to what extent the respondents use validation tools to check compliance of metadata, format or communication protocols and how complete they consider their metadata is. They also identified some barriers to metadata creation, like the complicated and contradictory guidelines. According to this survey, most of the respondents consider interoperability as a mainly technical issue.

3. METHOD

In this section, we outline the underlying method of our research. We aim to identify current and future main issues for the interoperability of web archives and digital libraries. We decided for an explorative, qualitative research in order to have the chance to identify novel issues in this field. Our intention was not to extract statistical results from either the entirety of the web archives and digital libraries or from a representative sample of it, but to gain useful insights from a group of people that are highly involved and particularly interested in this topic and the future progress. Hence, we chose the Delphi method to survey a purposive sample of experts.

The Delphi Method grew out of the need for a technique able to obtain the most reliable consensus of a group of experts [21]. While it was initially conceived as a group decision technique aiming to obtain a consensus, now it is also used as a research method to obtain reliable opinions and valuable contributions from a group of experts in order to resolve a complex problem [17]. For example, several Delphi studies are ranking-type and aim to extract a consensus opinion on the importance of specific issues, but others emphasise differences of opinion in order to develop a set of alternative future scenarios [21].

A Delphi method undergoes two or more rounds. The first round is an exploration of the subject. The researchers design the initial questionnaire and select an appropriate group of experts who are qualified to answer the questions. In this round each individual panellist contributes additional information that he feels is important to the topic [18]. The responses are then collected and analysed. Based on the analysed results, a second round is designed in which respondents are asked to revise their original responses and/or answer other questions based on group feedback from the first round. The Delphi method is an iterative process and each subsequent questionnaire is developed based on the results of the previous questionnaire. The number of the required rounds depends strongly on the purpose of the research. In general two or three iterations are suggested for most research but fewer could be also adequate to reveal sufficient information [24]. However, the participants are usually given at least one opportunity to revise their original answers upon examination of the group responses [18].

The Delphi study in our research consists of two rounds. A purposive sample of seven international experts from the web archiving and digital library communities was created. While the research team knew the identity of the participants, the participants were anonymous to each other. Thus, a possible bias by reputation or hierarchy perceptions or an answering according to expected norms could be avoided.

The aim of the first round was a brainstorming about the purposes, obstacles, possible solutions to overcome limitations, and other future challenges. Therefore, a questionnaire was created with four open questions (see Appendix A). Two researchers created the questions before an archivist reviewed them as domain expert. Based on the recommendations of the review, questions were adapted to improve the wording according to the participants' context. The final questionnaire was sent as text document and as online questionnaire to the participants at the beginning of February 2013. The participants had three weeks time to answer. Additionally, a reminder was sent in the middle of the three weeks to participants that had not responded yet. The final answers of the first round were analysed qualitatively by two researchers in parallel. Afterwards, results were compared and discrepancies in the interpretation were solved through discussion. The final results of the first round are presented in section 4 and were used to design the second round.

The aim of the second round was to verify identified results from the first round by all participants as well as to create further insights through evaluation regarding different aspects. Therefore, an online questionnaire with closed questions and the possibility for further comments was created. The questions were created by two researchers according to the structure of the first round's results, and reviewed afterwards by the archivist. Further improvements of the wording were made based on the review. Additionally, the questionnaire was tested with two individuals related to the archiving sector in order to test the understanding of the questionnaire as well as to confirm the time estimation for answering the questionnaire. The questionnaire was sent to the participants at the beginning of April, and a reminder was sent after two weeks to participants that had not responded yet. The second round was completed by six of the seven participants. The responses were analysed and the results are summarised in section 5.

4. RESULTS FROM THE FIRST ROUND

In the following, we present the results of the first round of our Delphi study. We structured the results into categories represented by the following subsections.

4.1 Purposes

We collect under the term purposes the motivations and abstract use cases that require interoperability. The identified purposes can therefore be understood as answers to the question why a web archive or a digital library would consider interoperability with other systems. In particular, the identified purposes can be overlapping or complementary and should not be understood as disjoint classes. However, each purpose may imply some specific requirements or a different context.

The identified purposes are further separated in three aspects. The first aspects describes the distinct uses for which interoperability is necessary. In this way, uses that motivate interoperability can be differentiated between (a) federated search, (b) federated access, (c) exchange, and (d) replication.

Federated search in the context of our research is the possibility to search from a single point or with a single query for data that are stored in several web archives or digital libraries. In the traditional library, for example, it enables the user to search various printed and electronic collections through one interface [6]. The search query that the user types in a single interface is sent to multiple search engines. In this way, it is common that a selection or subset of search engines is generated instead of broadcasting the query to all search engines. The typical phases of federated search are resource representation, resource ranking, distributed search, and result merging [5]. An example for federated search indicated by one of the participants was the following:

> *"For example, a collaborative of three of four cultural heritage institutions might digitize texts related to WWII and place them into a single collection. Each institution might house a copy of their own materials but create an aggregate index of all texts in the combined collection so that researchers may discover them and seek to access them from partner institutions as is feasible."*

While federated search requires that just the location of the desired objects can be found even if it is distributed in distinct archives, **federated access** also enables the user to retrieve the data directly from a single point. This means that the data can be, for example, viewed or downloaded. We distinguish between federated search and federated access in order to emphasise the opportunity for the user to directly access through one interface the objects that are

stored and managed in distributed locations. Therefore, a precondition of federated access is that the object has a digital form while federated search is also possible for non-digital, e.g. printed, objects. An example that indicated the desire for federate access was:

"One is to make it easier for people to access and use content despite the physical location of the content. For example a researcher can discover and bring together into one view content from many different repositories."

Exchange and replication are similar but describe different aims for the transfer of data between archives. The **exchange** of archived objects may be necessary to create or to complement specific collections like the collection of information about a specific topic or event. One participant reported:

"collaborative constitution of collections or exchange of collections between institutions. For example, constitution of web archives collections for the 2012 Olympic games in London (IIPC project)."[1]

Replication on the other hand aims at data redundancy in order to reduce the risk of data loss and improve reliability. The preservation of digital information has to consider physical threats (e.g. natural events, age of the hardware), technological threats (e.g. format obsolescence), human threats (e.g. curational errors), and institutional threats (e.g. economic failure). Replication combined with regular auditing can help to reduce the impact of these threats [2]. While the specific reasons for replication were not further explained by the participants, the need for replication was mentioned in statements like the following:

"The purpose of interoperability in the context of digital preservation is two-fold: exchange of information and distribution of replicas."

The second aspect derives from the differentiation in the scopes of the above uses. Hence, it can also be understood as a specialisation of the purposes already described. In particular, the following refinements were made about interoperation across:

- National boundaries,
- Organisational boundaries, either among organisations of the same type (e.g. among several digital libraries), or among organisations of different type (e.g. between a national digital library and the national web archive).

The last aspect that we identified differentiates the motivations based on the objects in focus. Thus, interoperability may concern either primary objects entirely or only metadata. One participant gave us the following example:

"It may be exchange of collection if data are interoperable, or only collaborative referencing of collections if only metadata are interoperable".

[1]For more information about the Olympics 2012 collection see also http://digital2.library.unt.edu/nomination/olympics2012/

4.2 Benefits through interoperability

Among the participants' views regarding interoperability, we identified also some benefits that arise from the institutions' interoperation and the general attempts in this direction. We consider as benefits any advantage or opportunity for the institutions and the involved communities that occurs through the interoperation of the systems or through the research and other efforts towards this. We distinguish the benefits from purposes since the latter are goals that we aim to achieve or problems that we try to overcome, while the benefits are the additional positive effects that arise through the process or the outcome. With respect to this, the following benefits were identified:

- Dissemination of the content of an institution's collections internationally. As stated by a representative of a digital archive which collaborates with a universal web archive organisation:

  "We are collaborating with X thanks to the presentation of our project on the website of X we can (get) not only a larger, but international attention."

- Institutions and organisations are benefited in areas in which they are constrained to act individually in terms of budget and annual resources or because of lack of know-how:

  "Creating interoperability requires more preparation and ongoing management but if executed well will result in benefits to an organization that could not be realized alone, especially in the domain of access or preservation, areas in which individual institutions are by nature constrained in terms of budget and resourcing on an annual basis"

  This point has been revealed as well in a previous survey [14] where respondents indicated a desire to engage in partnerships that could offer some technical assistance.

- Development of common tools to collect, exploit and preserve content:

  "Example : all IIPC members use the ARC or WARC standard so IIPC funds projects to develop or enhance ARC or WARC files harvesting, managing or accessing tools."

- Longevity of digital collections since their content is described and encoded in common standards. This particular point has been also investigated in [19] which examined digital longevity through standards and reached the conclusion that specific kinds of standards, even if not designed for digital longevity, are essential to this purpose to describe the functionality, the procedures and the concepts of a digital library or archive, to preserve the digital documents, to preserve the access to the content (metadata standards), and for interoperability.

4.3 Barriers to Interoperability

The second aspect of interoperability that we aimed to identify is the obstacles and limitations, or in other words, the barriers that hinder the establishment of interoperability. We grouped the identified barriers in five categories: (a) standardisation, (b) tools and implementation, (c) organisational obstacles, (d) legal problems, and (e) the approach to handle interoperability.

While various standards already exist, the current state of **standardisation** and compliance seems to be unsatisfactory. A lack of agreed standards has been reported. Similar to the lack of agreement, competition among the already existing standards has been reported.

However, even the agreements on standards do often not lead to interoperability because problems occur when they are applied or implemented. One problem is the **lack of tools** that implement the existing standards. Next to this, the same standard can be implemented differently in different contexts. More specifically, even if two archives apply the same schema, the content can be modelled differently and thus impede interoperability:

"Technically we model content differently. Even when we use the same schemas (e.g. METS) we use them in different ways."

While the barriers regarding standards are mainly of a technical nature, barriers occur also from an organisational and legal perspective. **Organisational obstacles** concern the ability and willingness of an organisation to provide interoperability for its collections. Some organisations are not willing to commit in collaborations and partnerships or they are not willing to invest in standardising processes:

"Too often organizations fear the process of becoming 'dependent on another organization' when it is hard enough to operate alone"

Furthermore, organisations may feel not able to provide or invest in interoperability because of the expected effort as well as the lack of know-how and resources in the organisation:

"Large-scale collaborations can be time-consuming and require a lot of effort and communication, especially for mission-critical activities like preservation."

Last, some organisations actually have no desire to provide any interoperability:

"In many cases, there is no desire for interoperability. Quite to the contrary, there are clear strategies aimed at not being interoperable in an attempt to lock in a user base, i.e. prevent users from seamlessly moving between information environments"

Legal barriers can hinder interoperability. Participants reported national regulations that limit or prevent any data exchange:

"exchange of data via ingest or export from other institutions outside of a 'national' umbrella is strictly limited or forbidden. This is true today for many EU countries like Denmark, Sweden and Norway"

This particular point has also been raised in previous survey [14] and was later addressed by the same author in detail [15]. Apart from this, the copyright holders define significantly the level of access and intellectual property laws hinder an open or public access:

"We rely on the personal permit of copyright holders. National libraries can't or do not offer free access to the collections."

Last, the **approach to establish or handle interoperability** seems to differ. For example, different perspectives between traditional librarians and web archivists were reported as a barrier to collaboration and interoperation between the two communities:

"there is sometimes a reluctance by the traditional library people to embrace web technology: harvesting and free text search versus a well controlled and high quality library catalog."

Furthermore, communities often define interoperability based on the specific systems they wish to interoperate and then define an approach to establish it, which is tailored to these systems:

"Often times, communities that are keen to achieve interoperability come at it from a perspective of determining which 'systems' need to be interoperable [...] This kind of system-to-system interoperability can effectively achieve desired interoperability levels among the targeted systems but leaves all other information environments unaffected and unable to benefit from the interoperability investment."

4.4 Suggested solutions & improvements

Several suggestions to overcome current barriers and achieve better levels of interoperability have been proposed by the participants as possible solutions or improvements.

Clear Legislation and policies regarding the exchange of data/metadata: An essential change would be clarity in national legislations regarding the exchange of data/metadata because it seems to be a grey area in many countries that makes the institutions more reluctant to exchange information.

"Today many believe a precedence has been set for this through the efforts of the Linked Open Data community (LOD) in Libraries, Archives, and Museums around the globe but in fact it is still a gray area in many countries making national institutions hesitant to exchange information regarding their holdings. With clarity on this front, the global archival community could work more closely and in partnership on capturing and preserving representative samples of the Web."

Standardisation: Regarding standards there seem to be a diversity of opinions. On the one hand, there is the belief that new, better, global and well-defined standards are needed, to handle interoperability limitations. For example, it should be very clear to institutions what is the minimum metadata information to be included in a single item:

"Defining a set of global standards and protocols for the exchange of this data will need to be ironed out including what minimal information must be contained in the core information package."

On the other hand, there is the belief that there is not really need for new standards, but there should be a consensus on which standards to use and then conformity with them. Furthermore, an initiative that would somehow necessitate the use of specific current standards would be beneficial.

Implementation & other developments: Even though the current standards seemed to be sufficient, the need for tools to implement them was also suggested:

"development of tools implementing current standards"

Further technical changes that are said to be supporting are the use of common APIs for search and retrieval and a central aggregation service that could bring all the information from several collections to the user. For example:

"we need to have common APIs for searching and retrieving content and metadata"

People's and communities' involvement: Communities and individual people are also said to play a part in this direction. The different communities should collaborate and be more involved in each other's activities so that their particular needs are also taken into account. For example, the web community could be more involved in the digital preservation community to ensure that web archiving needs are considered in the development of digital preservation standards:

"it is necessary to be involved in the wider digital preservation community in order to ensure that web archiving needs are taken into account by main digital preservation standards (eg METS or PREMIS)"

Involved people are also said to be influential because sometimes their community may significantly influence their perspectives. As mentioned previously, web and library world seem to have different and even controversial priorities sometimes and therefore, people with broader knowledge should be involved in the interoperability efforts:

"Different cultures: web people versus librarians. There are few people who belong to both worlds.[...]the most pressing need is the right kind of people. People who talk both languages."

Knowledge sharing is also another suggested important path. Sharing the experiences of various interoperability efforts, i.e. the successful stories, the failures and the practises that have been found to be best, would contribute to improve methods, avoid mistakes, and use resources more effectively. A consensus on the best practises and the sharing of them would contribute in more and more institutions joining and collaborating. This is not insignificant, since several institutions, especially libraries, don't have enough financial or personal resources to invest individually on such efforts. Therefore, an initiative or funded organisation to provide support about technical and legal issues would be also beneficial:

"As a institution financed by the university, public fundings and by projects we can't afford the costs for the technical support we need for the preservation. This means, we need an institution that helps with technical support. An EU-based organization that offers help for legal and technical questions"

Sharing knowledge should also include providing clear definitions and terminology about the digital preservation aspects.

Last, another recommendation suggests a different perspective, to consider **interoperability from the perspective of the web infrastructure** and implement it in terms of web and independently, creating information interoperability and diverge from system-based interoperability:

"tackle interoperability not from a repository, digital library perspective but rather from the perspective of the web infrastructure. Assets in archives and digital libraries are web resources with URIs. If interoperability for such assets is required, define and implement it in terms of the web."

4.5 Interoperability perspectives

The responses of the first round revealed another dimension of interoperability based on the perspective that is considered. From this point, two different perspectives can be distinguished:

System Interoperability (or system-to-system interoperability) that is probably the most traditional and common perspective which communities tend to follow. It is the perspective of defining interoperability based on which systems are desired to interoperate. This perspective might be quite successful but it is limited to the particular targeted systems:

"This kind of system-to-system interoperability can effectively achieve desired interoperability levels among the targeted systems but leaves all other information environments unaffected and unable to benefit from the interoperability investment."

Information interoperability is about putting the focus on the information itself and making the information interoperable with different systems. It is the perspective of considering interoperability not from the perspective of a digital library, repository or any other information environment but rather from the perspective of web infrastructure instead:

"An approach that yields better return on investment is based on achieving the desired level of interoperability by specifying and implementing it in terms of the existing infrastructure (the Web and its fundamental building blocks): define the interoperability problem in terms of the web and its primitives and solve it using those primitives, web standards, widely embraced technologies. [...] Assets in archives and digital libraries are web resources with URIs. If interoperability for such assets is required, define and implement it in terms of the web."

4.6 Further challenges

Part of our research was to examine interoperability with a view in the future. Therefore the participants were asked about future challenges they consider. We include in this category either the forthcoming changes that will put additional difficulties to interoperations or the challenging goals that have to be considered in further steps. With respect to this, four future challenges have been identified. It should be noted in advance that not all of them are directly related to interoperability, but primarily related to web archiving issues. They are stated, nonetheless, on the one hand because

the interoperability of web archives is significantly dependent on web archiving strategies, and, on the other hand, to support further web archiving discussions and developments.

Interoperability of the content: While current efforts aim on the interoperability of the systems to enable search, access, and transfer of resources, future attempts will focus also on the interaction of content. The vision could be a seamless web of archived content.

"The most immediate challenge I see is the need/desire to start looking at web archives and digital libraries not only as a collection of resources with URIs but also as big datasets. This means that, not only will it be important to be able to have interoperability expressed in terms of URIs, metadata but also in terms of content."

New players with different systems, needs, and tools are emerging in the field of web archiving:

"However, new actors are emerging, eg research labs or private companies that may use specific tools and/or are not experienced with the necessity of respecting standards. [...] So there is a strong need: - to promote standards towards new actors in web archiving"

The increasing efforts to archive as much of the web as possible combined with the immense growth of the web will lead to an explosion of the **amount of web data** to archive:

"Furthermore, the volume of data has exploded to 500TBs to PBs of data per crawl of the Web."

New and complex media and web resources (Web 2.0, Social Media, etc.) demand enhanced methods for web preservation. For example:

"The problem of preserving social networks. For example, Facebook is, for the moment, a very important communication tool in the literary field, but because of the legal obstacles it is impossible to archive Facebook-pages (it would be only possible, if it would be possible to cut all comments and posts from other authors than the rightholder)."

5. RESULTS FROM THE SECOND ROUND

In the following, we summarise the results of the second round of our Delphi study. In this round, each panellist received the group response, structured as closed type questions, and was asked to evaluate it. Therefore, participants had the chance, on the one hand, to revise or confirm their own original answers, and on the other hand to read and consider the other panellists' views. They were also given the option to add comments and, therefore, the chance to object, clarify, complete the existing statements or add a new one. Due to constraints of questionnaire research, we focussed the second round on the five core aspects: purposes, barriers, suggested solutions, further challenges, and perspectives of considering and realising interoperability in web archives and digital libraries.

We asked for agreement regarding the identified **purposes** on a four point Likert scale from "Strongly disagree" to "Strongly agree". Each of the identified purposes was agreed by at least five of the six participants. Two times, a participant answered with "I can't say", and one participant disagreed on replication as a purpose. In summary, we assess the purposes as verified. Minor trends can be identified in the differences of strong and normal agreement. Federated access and federated search got stronger agreement than exchange and replication. Also the interoperability of metadata got stronger agreement than primary objects.

The **barriers** were evaluated with four point Likert scale from "Not a barrier" to "Extreme barrier". Additionally, the participants were asked to evaluate them separately from the point of view of an individual organisation (e.g. a single library) and of the community as a whole. Verification of an identified barrier had to be negated if it was assessed as "Not a barrier" for both cases of single organisation and entire community by at least one participant. Based on the results, all identified barriers were verified except the *competition among current standards*, and the *unwillingness of institutions to invest in standardising*. Furthermore, we evaluated the consistency of the group responses through analysing the standard deviation for the verified barriers. Thus, we can estimate the agreement among the participants for each barrier. The responses were most consistent for the barriers of *lack of resources (in the organisation)*, and *different perspectives and priorities between different communities*. The least agreement among the participants existed for a *lack of agreed standards*, and the barrier of *locked systems & no desire for interoperability*. In general, the responses for the community perspective were more consistent than for the view of a single organisation. Furthermore, the impact of the barriers was in most cases higher for the community perspective than for the organisation's view. The strongest barriers from the view of a single organisation are the *lack of resources (in the organisation)*, *different implementations of the same standard*, and *intellectual property laws*. The strongest barriers on the community level are *limited or forbidden exchange of data outside national borders*, *lack of resources (in the organisation)* and *intellectual property laws*.

Next to verifying the **suggested solutions** to overcome existing barriers, they should also be assessed regarding their efficiency in order to deduce recommendations which solutions may be prioritised. Efficiency consists of the ratio of the impact of the solution to the effort to realise the solution. The impact was measured through the evaluation of effectiveness on a four point Likert scale from "Not effective / Not a solution" to "Very effective". The effort was measured through difficulty on a five point Likert scale from "Very easy" to "Very difficult". All the identified suggestions for solutions were verified. One participant rated *consensus on current standards and conformity* and *foundation of central organisation that provides support for technical & legal issues* with "Not effective / Not a solution" but did not assess it as "Not a solution" in the difficulty measure. Therefore, we deduce that he assessed the solution as not effective but verified it as a generally possible solution. In order to provide recommendations about the identified solutions, the average effectiveness and average difficulty for each solution were calculated and plotted in a portfolio (see figure 1). Three clusters can be identified:

1. **Highly recommended solutions**: A line was drawn from "somewhat effective" and "very easy" to "very effective" and "difficult". Solutions in the sector above the line are considered as very efficient because their estimated effectiveness is higher than the required effort to accomplish. The most promising solution is

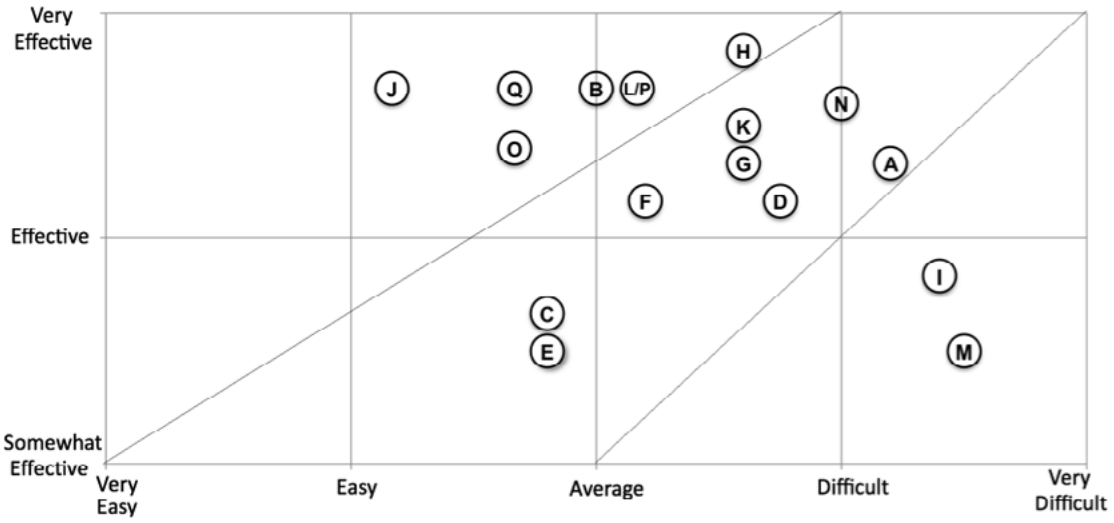


Figure 1: Portfolio of suggested solutions: (A) Consensus on current standards and conformity with them; (B) Initiatives / projects to necessitate the use of current standards; (C) Enhancement of current standards; (D) Global & well-defined standards; (E) Development of new standards; (F) Promotion of current and new standards; (G) Development of tools that implement standards; (H) Common APIs for search & retrieval; (I) Central aggregation service; (J) Sharing experiences, best practices & successful stories; (K) Consensus on best practices; (L) Clear definitions & terminology about digital preservation; (M) Foundation of central organisation that provides support for technical & legal issues; (N) Clear legislation & policies for the exchange of data / metadata; (O) Define interoperability from a Web infrastructure perspective instead of a system-to-system perspective; (P) Better collaboration and stronger involvement of related communities to each other's activities; (Q) Involvement of people with broader knowledge / experience, not individually confined to community aspects

thereby *sharing experiences, best practices & successful stories*.

2. **Recommended solutions**: A line was drawn from "Somewhat effective" and "average difficult" to "very effective" and "very difficult". Solutions that lie above this line and below the sector of highly recommended solutions can be assessed as efficient because their effectiveness still justifies their effort. It is notable that most of the solutions that are related to standards are located in this sector (A, C, D, E, F, and G).

3. **Inefficient solutions**: Solutions in the third sector can not be assessed as efficient because their effectiveness is much lower than the probable effort to realise them. With a *central aggregation service* and a *foundation of central organisation that provides support for technical & legal issues*, it is striking that the only two solutions that suggest a centralised service or institution are located in this sector.

Further challenges for the future were evaluated on a four point Likert scale from "Not a priority" to "High priority" and the alternative option of "Not a challenge". Each challenge was rated at least with "Low priority" by all participants, and, thus, the four identified challenges were verified. The average priority of each challenge was above medium priority. The increasing complexity of web resources was considered as the most pressing challenge.

The last part of the second round's questionnaire aimed at a comparative evaluation of system-to-system interoperability and information interoperability. None of the participants questioned the general applicability of the perspectives, and, therefore, it can be considered as verified by the participants that both are possible ways to establish interoperability. However, the answers to the comparative part were quite heterogeneous, and do not allow the identification of a clear trend.

6. DISCUSSION

In this section, we discuss the results of our Delphi study. As a first result we identified several purposes or use cases that demand interoperability. The reasons for interoperation of web archives and digital libraries can be generalised into two aims. On the one hand, the user should be able to have access to collections or individual resources that are archived in one or more distinct repositories regardless of their location. This can be carried out by federated search, federated access, and through the exchange of objects in order to create a new collection. On the other hand, interoperation is required to establish the replication of objects into different locations, and, thus, reduce the risk of loss caused by several threats [2]. However, the identified purposes of interoperability were not as manifold as we expected. For example, interoperation that is necessary for sophisticated analysis on web archives, e.g. link analysis [22], as well as any interoperation demands for the ingest of new digital content into a web archive or digital library has not appeared in the participants' statements.

Additionally, we identified several benefits that are con-

nected to interoperability. Thereby, the interdependence between collaboration and interoperability become apparent. For example, the common agreement on specific standards for interoperation facilitates collaborative efforts for the development of tools as well as the knowledge exchange regarding common problems. This in turn facilitates higher levels of interoperability.

The identified barriers and solutions are connected by nature because a solution (or improvement) addresses one or more barriers. Therefore, the categories we identified are also similar for both. However, when we compare the identified barriers and solutions with the existing interoperability models from the beginning of this paper, two peculiarities have to be noticed. Firstly, perspectives that include also higher levels, e.g. the organisational level, seem to be more appropriate to consider interoperability for web archives and digital libraries. Thereby, a lot of problems on lower level can be addressed through further standardisation efforts while this is hardly possible on higher levels, e.g. the lack of knowledge or fears in the organisation. Secondly, a perspective or level that focuses on legal issues is not mentioned explicitly in the presented models while it can be highly restrictive for interoperability attempts. Therefore, existing models for interoperability should be adapted in order to emphasise the importance of legal considerations, especially in the domain of web archives and digital libraries.

Another important finding is the identification of different ways to understand interoperability, and, thus, to establish the interoperation between different systems. Interoperability is most commonly considered as a task between two systems where both can take specific roles, for example a provider and a consumer of data [4]. Thus, the requirements are derived from the interoperation task and the systems characteristics, and the interoperability may be specifically adjusted to the corresponding systems even if the use of standards facilitates the same or similar interoperation with other systems. Contrary, the perspective of information interoperability abstracts from the specific systems, and aims on the provision of data as entities that support undetermined uses. Therefore, the entity must comprise or link all information necessary for processing in an undefined scenario.

In the second round of the study, almost all the results from the first round were verified and the evaluation allows further findings: Federated search and federated access together with the exchange of the metadata seem to be more present as interoperability purposes than the replication and the exchange of primary objects. The barriers that hinder or prevent interoperability are manifold. The most salient are the lack of resources to establish interoperability, different implementations of standards even if the same standard is used, intellectual property laws and limited or forbidden exchange of data outside national borders. They show that interoperability is dependent on organisational, legal, and technical aspects with little or no indication that one aspect may be more important than the other. The evaluation of suggested solutions revealed that the most promising are these that comprise involvement or knowledge sharing of the community like sharing experiences, best practices & successful stories, involvement of people with broader knowledge & experience, clear definitions & terminology, and better collaboration and stronger involvement of related communities to each other's activities. On a lower level but still recommendable is the majority of solutions that are related to standards and tool development. However, the creation of centralised services or support institutions can be hard to recommend because the estimated impact does not legitimate the expected effort.

7. CONCLUSION

The Delphi study, presented in this paper, revealed insights regarding current problems, limitations, needs and challenges that are encountered in today's interoperations (or efforts in this direction) among systems of the web archiving and digital library communities. The study was carried out among a small, purposively selected group of people with expertise on the topic, who shared their views and ideas, adding a valuable input to the research. It offered a unique contribution to the research field of interoperability, presenting the current barriers but also suggestions for future approaches, and can be a useful study for the communities of web archiving, digital libraries, and digital preservation.

However, a limitation has to be taken into consideration. The findings are influenced by the selection of experts. Therefore, the same questions may lead to different results with other experts. However, we did not aim on completeness, and we consider it unlikely that such results would be conflicting.

Finally, it should be emphasised again that further studies should be conducted in order to validate and to extend the understanding of current and future interoperability aspects for web archives and digital libraries.

8. ACKNOWLEDGMENTS

This work was conducted as part of the BlogForever[2] project co-funded by the European Commission Framework Programme 7 (FP7), grant agreement No.269963.

9. REFERENCES

[1] Web Archiving in Europe. Technical report, Internet Memory Foundation, 2010.

[2] M. Altman, M. O. Adams, J. Crabtree, D. Donakowski, M. Maynard, A. Pienta, and C. H. Young. Digital Preservation through Archival Collaboration: The Data Preservation Alliance for the Social Sciences. *The American Archivist*, 72(1):170–184, 2009.

[3] W. Y. Arms, D. Hillmann, C. Lagoze, D. Krafft, R. Marisa, J. Saylor, C. Terrizzi, and H. Van de Sompel. A Spectrum of Interoperability. *D-Lib Magazine*, 8(1), Jan. 2002.

[4] G. Athanasopoulos, L. Candela, D. Castelli, K. El Raheb, P. Innocenti, Y. Ioannidis, A. Katifori, A. Nika, S. a. Ross, A. Tani, C. Thanos, E. Toli, and G. Vullo. Digital Library Technology & Methodology Cookbook: Interoperability Framework, Best Practices & Solutions. Technical report, 2011.

[5] T. T. Avrahami, L. Yau, L. Si, and J. Callan. The FedLemur project: Federated search in the real world. *Journal of the American Society for Information Science and Technology*, 57(3):347–358, 2006.

[6] A. Curtis and D. G. Dorner. Why Federated Search? *Knowledge Quest*, 33(3):35–37, 2005.

[2]http://blogforever.eu/

[7] A. Geraci, K. Freny, L. McMonegal, B. Meyer, J. Lane, P. Wilson, J. Radatz, M. Yee, H. Porteous, and F. Springsteel. *IEEE Standard Computer Dictionary: Compilation of IEEE Standard Computer Glossaries*. IEEE Press, Piscataway, NJ, USA, 1991.

[8] D. Gomes, J. Miranda, and M. Costa. A survey on web archiving initiatives. In S. Gradmann, F. Borri, C. Meghini, and H. Schuldt, editors, *Research and Advanced Technology for Digital Libraries*, volume 6966 of *Lecture Notes in Computer Science*, pages 408–420. Springer Berlin / Heidelberg, 2011.

[9] P. Gottschalk. Maturity levels for interoperability in digital government. *Government Information Quarterly*, 26(1):75–81, Jan. 2009.

[10] S. Gradmann. Interoperability. A key concept for large scale, persistent digitallibraries. Technical report, 2007.

[11] A. Grotke. International Internet Preservation Consortium: 2008 Member Profile Survey Results. Technical report, 2008.

[12] IDABC. European Interoperability Framework for pan-European eGovernment Services. Technical report, Luxembourg, 2004.

[13] P. Innocenti, M. Smith, K. Ashley, S. Ross, A. De Robbio, H. Pfeiffenberger, and J. Faundeen. Towards a Holistic Approach to Policy Interoperability. *The International Journal of Digital Curation*, 6(1):111–124, 2011.

[14] G. Jacobsen. Webarchiving Internationally: Interoperability in the Future? In *World Library and Information Congress: 73rd IFLA General Conference and Council*, Durban, South Africa, 2007.

[15] G. Jacobsen. Web Archiving: Issues and Problems in Collection Building and Access. *LIBER Quarterly*, 18(3/4):366–376, 2008.

[16] C. Lagoze, H. Van de Sompel, M. Nelson, S. Warner, R. Sanderson, and P. Johnston. A Web-based resource model for scholarship 2.0: object reuse & exchange. *Concurrency and Computation: Practice and Experience*, 24(18):2221–2240, June 2010.

[17] J. Landeta. Current validity of the Delphi method in social sciences. *Technological Forecasting and Social Change*, 73(5):467–482, June 2006.

[18] H. A. Linstone and M. Turoff. *The Delphi Method: Techniques and Applications*. Addison Wesley, 2002.

[19] H. H. J. Lorist and K. v. d. Meer. Standards for Digital Libraries and Archives: Digital Longevity. In *NDDL '01 Proceedings of the 1st International Workshop on New Developments in Digital Libraries: n conjunction with ICEIS 2001*, pages 89–98. ICEIS Press, 2001.

[20] National Diet Library. CDNLAO Questionnaire Survey on Web Archiving. Technical report, 2010.

[21] C. Okoli and S. D. Pawlowski. The Delphi method as a research tool: an example, design considerations and applications. *Information & Management*, 42(1):15–29, Dec. 2004.

[22] E. Reynolds. Web Archiving Use Cases. Technical report, 2013.

[23] R. Shiozaki and T. Eisenschitz. Role and justification of web archiving by national libraries: A questionnaire survey. *Journal of Librarianship and Information Science*, 41(2):90–107, 2009.

[24] G. J. Skulmoski, F. T. Hartman, and J. Krahn. The Delphi Method for Graduate Research. *Journal of Information Technology Education*, 6, 2007.

[25] A. Tolk, S. Y. Diallo, and C. D. Turnitsa. Applying the Levels of Conceptual Interoperability Model in Support of Integratability, Interoperability, and Composability for System-of-Systems Engineering. *Journal of Systemics, Cybernetics and Informatics*, 5(5):65–74, 2007.

[26] G. Vullo, G. Clavel, N. Ferro, S. Higgins, R. van Horik, W. Horstmann, and S. Kapidakis. Quality interoperability within digital libraries: the DL.org perspective. In *2nd DL.org Workshop in conjunction with ECDL 2010*, Glasgow, UK, 2010.

APPENDIX

A. FIRST ROUND'S QUESTIONS

All questions of the first round were formulated as open questions.

- What in your view are the purposes of interoperability? What problems or opportunities are addressed with interoperability? Please reply with a descriptive answer, if possible using scenarios that describe the purpose, the partner institutions, and the systems that are involved.
  Think of problems that have been solved or problems that exist and require interoperability practices, problems that you either experience directly or you can identify. Additionally, think of benefits that occur from the interoperation between systems/institutions.
- What are the main obstacles and limitations that prevent or hinder interoperability?
  (technical, political, organizational, management, legislation or other barriers)
- What changes or developments in the landscape would, in your view, assist the interoperability of digital libraries and/or web archives (and how)?
  Think of technical changes/developments (e.g. standards, frameworks, services), political or legislation changes, new concepts etc.
- What do you consider as future challenges regarding interoperability of digital libraries and/or web archives?
  Think about important problems that have to be solved, obstacles to overcome, possible additional future barriers that may occur due to forthcoming changes in needs, technology, perspectives, legislation etc.

Database Preservation Evaluation Report - SIARD vs. CHRONOS

Preserving complex structures as databases through a record centric approach?

Andrew Lindley
AIT - Austrian Institute of Technology GmbH
Donau-City-Strasse 1, 1220 Vienna, Austria
andrew.lindley@ait.ac.at

ABSTRACT

Preserving information systems is one of the greatest challenges in digital preservation. In this paper we outline the existing strengths and shortcomings of a record-centric driven preservation approach for relational databases by lining up a state-of-the art industry database archiving tool CHRONOS[1] against SIARD[2] one of the most popular products in the GLAM (galleries libraries archives museums) world. A functional comparison of both software products in the use cases of database retirement, continuous and partial archiving as well as application retirement is presented. The work focuses on a technical evaluation of the software products - organizational and process aspects of digital preservation are out of scope. We explain why preserving complex structures as databases through a record centric approach does not only depend on the amount of information captured in the preservation package and present a brief overview on available functional aspects in CHRONOS that help to address the challenges of application decommissioning. The paper at hand presents the results of a case study which was undertaken 2012 at AIT - Austrian Institute of Technology GmbH.

Categories and Subject Descriptors

H.2.m [**Database Management**]: Database Applications—*Miscellaneous*; D.3.4 [**Information Storage and Retrieval**]: Systems and Software—*Performance evaluation (efficiency and effectiveness)*; H.3.7 [**Information Storage and Retrieval**]: Digital Libraries—*System issues, User issues*

General Terms

Verification, Experiment, Performance, Reliability, Management, Human Factors

[1] `http://www.csp-sw.de`
[2] `http://www.bar.admin.ch/`

Keywords

Digital Preservation, Database Archiving, Case Study, Technical Evaluation, Decommissioning, Application Retirement

1. INTRODUCTION

Sustained information to our scientific and cultural heritage world is stored digitally. The term digital preservation (DP) summarizes methods and techniques to secure long-term access to digital information. Every information management system, data warehouse, or even the simplest online webstore is backed by a database system. For the last decades relational databases have been the dominant technology in this area mainly due to broad vendor adoption and acceptance of the SQL standard for the relational model. ACID (Atomicity, Consistency, Isolation, Durability) provides principals governing how changes are applied to a database. In the decade of big data some of these principles are loosened with respect to high data volumes and high traffic throughput and niche products as NoSQL databases, key value and tripple stores found their place.[1]

Within the last ten years the digital preservation community was able to achieve a solid understanding of issues and provided solutions and guidance in the domain of document preservation. The currently ongoing European initiatives widen the domain of digital preservation taking on from memory institutions and include scenarios such as healthcare, data with direct commercial value and web-based data and focus on aspects such as data collection, scalability, reconfigurability and lifecycle management. [2], [3]

Preserving information systems is one of the greatest challenges in digital preservation. The paper at hand presents the results of case study which was undertaken 2012 at AIT. The technical evaluation and comparison of the database preservation tools SIARD [4] and CHRONOS [5] targeted at the use cases of database retirement, partial and continuous archiving as well as application retirement. Besides presenting a functional tool comparison we highlight strengths, shortcomings and white spots in general. A goal is to broaden the discussion on database preservation by comparing one of the most popular tools for database preservation in the GLAM domain against CHRONOS a leading industry product. This work focuses on a technical evaluation of the software products and only briefly covers organizational and process oriented aspects of digital preservation. CHRONOS is a commercial product owned by CSP

and emerged through a joint research cooperation between the department of computer science at the university for applied science in Landshut. SIARD (Software Independent Archiving of Relational Databases) is owned by the Swiss National Archives (BAR) and is both an open format to express relational database archives as well as a software product SIARD suite. It is available under closed-source license and was originally developed by Trivalis.

1.1 Continuous Archiving and Application Retirement

Solutions for database archiving are not part of a standard relational database systems. According to Forrester [3] only 15% percent of business data are actively required to serve a company's day-to-day requirements while the vast amount of data could already be moved into an archived state.

> [..]Terabyte-size transactional databases are harder to manage, increase costs for hardware capacity and database licenses, and drive up requirements for database administrators (DBAs). Yet 85% of production data is inactive, so information and knowledge management professionals should devise a database archiving strategy that moves inactive data to lower-cost storage and servers, thus improving the manageability, performance, and security of critical production applications[..]

A typical data life-cycle can be categorized in an

a) Active State, in which data is generated and modified as part of the production system

b) Archiving State, in which a dataset is no longer altered but still needs to be kept active for fulfilling existing business processes

c) Long-Term Archiving state, in which only selected parts of a dataset are kept for retention

Effects that are achievable with a continuous database archiving strategy are for example the reduction in database license fees, easier adherence to SLAs, efficient system consolidation or a noticable reduction of maintenance effort. Such a system can also be constructed to adhere to different requirements as for example given legal regulations on data retention.

Preserving complex structures as databases through a record centric approach does not solely depend on the amount of information captured in the preservation package but requires a surrounding process to capture all required metadata as additional documentation and understanding of the underlying data. This is shown by a case study of the National Archives of the Netherlands [6] in 2011 on longterm preservation of relational database systems in coherence with the legal mandate to archive public data records (content, reports, applications) and records from public institutions. They came to the insight that even though an acceptable number of available tools and technology was available to address the challenge, there was a lack on sufficient knowledge on the relevance of the archived data and its contextual relationship within given business processes.

Decommissioning is the process of a planned shut-down and removal from operational use. Decommissioning as well as application retirement are challenges that an archive or library is confronted with. In the ECM podcast [7] on practical digital preservation Adrian Brown, director of the Parliamentary Archives in London mentions the 'blurring of the boundary' between digital objects and the applications that they are held in as key challenge the institutions are confronted with. Digital preservation initiatives and projects made great progress in tackling the problem of how to preserve the file formats and the objects themselves but now faces the more complex problem of how to preserve the information that an application has about the objects it holds? How to enable digital objects to move from one application to another without losing that information? Within this paper we present technical issues and generic challenges we discovered when transforming a database into a long-term database archive by using the tools SIARD and CHRONOS and conclude with a brief overview on the features that CHRONOS is able to deliver for the application retirement scenario.

1.2 Paper Outline

In the first part of the paper we raise relevant research questions and point out existing limitations of a record-centric driven preservation approach for relational databases. In the second part we present the experiment setup and detailed evaluation results in a functional comparison. We conclude with a brief overview on functional aspects provided by CHRONOS for the challenges of application decommissioning.

1.3 Related Work

Burda et al.[8] present a semantic literature review of 122 publications in the domain of digital preservation with respect to different aspects as drivers, stakeholders and applied research methods in the field. The authors disclose the gap of a DP reference model that addresses organizational concerns, considering aspects such as costs, risks, decision criteria, etc. The ISO standard 'Reference Model for an Open Archival Information System' (OAIS) which guarantees cross-organisational concepts and terminology has impact in the construction of a preservation package and the 'Model Requirements Specification for the Management of Electronic Records' (MoReq2) which provides principles to guide institutions in the implementation of electronic record management systems are both seen as relevant in the domain of database preservation. Preservation Planning Tools as PLATO [9] support the process of cost-benefit analysis within digital preservation decision making but to the author's knowledge no case study on database preservation was conducted to date. Digital preservation projects co-funded by the European Commission under the sixth and seventh framework programs are given in [10] which presents objectives, developments and major outcomes of the projects. The intellectual property rights of SIARD lays at the Swiss Federal Archives and development was stimulated through

[3]http://www.forrester.com/Database+Archiving

the Planets project [11]. A different approach than extracting and describing relational data through generic and vendor independent XML formats as DMBL[12] or SIARD for archival and cross compatibility purposes is the preservation of relational data through RDF triples as implemented in the Semantic Archive and Query (SAQ) system where access is provided via A-SPARQL queries.[13]

Preservation of databases and database records has always been an an important task for national archives which in many cases is based on a legal mandate to preserve governmental records. Activities in this area for the Danish National Archives started in 1973. In 2008 all of the approximately 3.600 Archival Information Packages (AIPs) held in their collection were exports from database systems, whereby content from both business systems and record management systems are transferred as relational databases. The Access project was completed 2008 and since September 2010 archival records, which are structured according to the Danish archival standard for digital records are delivered in a modified version of the SIARD format which also includes contextual documentation. A general query building system for archival records has been developed to support unknown needs for retrieving data. [14]

2. EXPERIMENT SETUP

Work presented in this paper is based on a case study which was undertaken by AIT in 2012. The report is split into three major sections, a generic evaluation of the underlying tools and their technical features, a ISO 25010:2011 driven evaluation of software quality aspects based on ISO/IEC TR9126 'quality in use' metrics in the areas of efficiency, productivity, security and satisfaction within a very specific usage context and staging environment, and finally an interpretation of the research results based on the customer's requirements. Please note that part two and three of the report itself are confidential as they contain customer sensitive information and therefore are not presented in this paper. Aspects as licensing or pricing information from part one of the report which are protected by NDA agreements are also left out.

Database preservation strategies heavily depend on the nature of the underlying data where typically three main categories are distinguished: administrative, scientific and document management databases[15]. Part one of the tests which are presented within this paper were executed on a virtualized standard desktop hardware infrastructure running Windows-7 with a local copy of the tools and all required software dependencies together with an Oracle 11gR2 database filled with Transaction Processing Performance Council (TPC)-C "'Entry-Order"' records that were enriched with BLOB and CLOB data. The aim is not to provide benchmark information but rather accompanying documentation on technical features and unique selling points - no entitlement of functional completeness.

3. EVALUATION RESULTS

A quick overview of the product driven evaluation results is given in Table 1. More detailed explanations on the individual items and resulting issues are given within this paper.

Evaluated Categories	Siard	Chronos
Supported Preservation Scenarios	3/10	8/10
Exported Elements of an Archived Database	6/10	8/10
Pre- and Postprocessing via Database Scripts and Markertables	5/10	10/10
Data Retention and Data Controls	1/10	10/10
Support of UDTs and Oracle Specifics	3/10	5/10
Rights, Roles and User Management	0/10	9/10
Archive Data Access and Performance	2/10	10/10
Syntactic and Semantic Data Changes	0/10	9/10
Existing APIs and Interfaces	3/10	8/10
Scalability and Limitations	7/10	9/10
Risk Behavior and Dependencies	9/10	8/10
Referential Dependencies	3/10	10/10
Standard and Compliance	4/10	4/10
Data Exchange Formats	5/10	5/10
Structure, Setup and Size of the physical Archive	7/10	7/10
Specification of Information Lost	3/10	3/10
Installation and Delivered Components	10/10	9/10
License Models, Costs and Reference Customers	5/10	5/10

Table 1: Overview of the Product Driven Evaluation Results

Supported Preservation Scenarios

The evaluation is based on the support of the three classification scenarios: 'database retirement', 'ongoing or partial database archiving' and 'application retirement'. Questions addressed are to which degree do the products offer support for database retirement (including database independent transformation, understandability of the physical archive, SQL data access, etc.), continuous or partial archiving (inc. schema changes over time, data retention, etc.) and application retirement (incl. available support for the recreation of business objects, functions as reporting, data access roles and programmatic access, etc.).

CHRONOS is able to deliver an extensive package with support for all three database preservation scenarios. Especially 'database retirement' and 'continuous/partial archiving' are seen as core use cases which are covered out of the box in the requested and required complexity. A key feature of CHRONOS regarding data access is the possibility to execute SQL92 compliant reporting through queries on top of the archived datasets. Even though the content is exported and physically stored in basic text files the query performance is comparable to the one of a relational database. The scenario of 'application retirement' is backed through the Chronos software module Archive Explorer that allows recreating relevant business objects, custom views and reporting workflows based on the archival records. All modules adhere to data access and role policy models. CHRONOS software suite can in addition leverage positive secondary effects as quicker backup and restoration time, as an easy way of generating snapshot data, performance improvements within the production database and reduction of storage and licensing costs as typically database system are licensed by number of cores.

SIARD is defined to fully support the 'database retirement' use case for a huge number of relational database systems. Support for the scenarios application retirement or continuous / partial archiving are not envisioned for the SIARD Suite. Even though it is possible to re-import a SIARD database archive by restoring its primary tabular data into a RDBMS in order to execute complex queries and even though it is possible to manually rebuild or ignore the lost metadata such as views, procedures, triggers, etc., the system it is not meant to re-vive a database for continuously exporting data.

Exported Elements of an Archived Database

A relational database and RDBMS is a complex product that technically speaking consists out of Tables, Views, Materialized Views, Indices, Packages, Triggers, Stored Procedures, Functions, Sequences, Scheduler, Check Constraints and Triggers, Queues, Database Links, User Management Access Privileges and Roles to mention the most important constructs. Which database elements are extracted into a database archive by the preservation tools at hand? Which of these elements remain functional after re-importing them into a RDBMS and which of them solely serve the purpose of documentation within an archive?

The main focus in CHRONOS lays on exporting primary data and datatypes. Tables, Views, Indices, Packages, Procedures, Functions, Triggers, Sequences, Materialized Views, Scheduler and Check Constraints are supported elements when transferring data into a database archive. Queues are not preserved as they only serve for communication purposes and no value is seen in keeping them. Database Links are not supported. Jobs are deprecated and are not archived by CHRONOS. *User management and definition of roles are not preserved by CHRONOS as there is no access mechanism through standard interfaces.* In many cases user and rights management however is not depicted at database level anyway. On a functional level CHRONOS offers extensive support for integrating with central policy and access permission systems as LDAP. *Triggers, Procedures and Views are exported from the production system but remain unsupported elements when re-importing the archived data into a RDBMS.* This can be seen as a security feature as cross mapping between different database vendors and also between versions of the same product (e.g. Oracle version 10 and 11) would have the potential to cause serious inconsistencies.

SIARD exclusively supports archiving of core SQL:1999 elements. Procedures and Functions are minimally supported and documented in a SIARD archive, depending on accessibility of the pertinent metadata information. The tool does not support functional long-term preservation of code but concentrates rather on preserving primary data. Triggers are supported by the SIARD format as they are defined in SQL:1999 but are not archived by SIARD Suite as they are only seen useful for 'live' databases where activities occur that trigger them. Materialized Views are not defined in SQL:1999 and in most database systems they are just (temporary) tables. Check Constraints are supported by the SIARD format as they are defined in SQL:1999 but usually are not archived as they are not easily accessible in most database systems. Users and Roles are archived by SIARD Suite. Standard 'scalar' SQL data types (Strings, Numbers, Dates) are supported by SIARD. User-defined data types (UDTs) at the moment are not archived, because no real life database system supported them when the design and development of SIARD started. There are plans to enhance the SIARD format to accommodate UDTs, however backward compatibility between the different versions of the SIARD format is a major requirement! Database links and packages are not supported. Packages are not defined in SQL:1999 and are not supported by all relational database systems. Indices are not supported, as indices in SQL:1999 are not defined as database elements but only serve as performance enhancers. Also Queues and Sequences are neither defined in SQL:1999 nor supported by all relational database systems or SIARD. *When re-importing a SIARD archive into a RDBMS, solely tables and tabular content is restored. Constraints are attempted to be restored as primary and foreign keys. Views, procedures, users, triggers, and check constraints are not restored as they could cause problems between different database instances.* From a SIARD perspective views, procedures, triggers, etc. are just considered as metadata. This information is therefore only depicted within the metadata.xml file, which is located in the header and not in the content folder for primary data. SIARD concentrates on restoring primary table data in RDBMS for the purpose of executing complex queries on it.

Pre- and Postprocessing via Database Scripts and Markertables

In the process of creating an archival package, especially in the scenario of partial and ongoing archiving, it might be necessary to execute pre- and post processing steps on the database as for example preparation or cleanup tasks. Supporting a smooth and integrated continuous archival workflow might require logging some kind of state or placings process markers within a production system. To which degree do the tools offer support for interacting with a production environment as executing pre- or post processing scripts or documenting archival state within the database itself?

CHRONOS allows to directly interact with a database system via shell commands, database scripts and marker tables. Documentation within a database system is possible via marker tables at the granularity of individual records. *SIARD, by design, never writes to a database and can therefore be executed with read-only permissions.* In the SIARD archive date and the circumstances of the download are recorded. SIARD Suite does not directly support pre- or post-processing of database scripts as this is highly dependent on the database system in use. Due to the fact that the SIARD Suite not only provides a GUI application but also supports the command-line interface for up- and download of archives, there is a workaround for calling a script or batch file via sqlplus for static pre- and post processing.

Data Retention and Data Controls

Due to legal regulations for example on handling of personal or sensitive data it might be required to keep and/or delete records after a given period of time from an archive. Other forms of data retention concern the periodical refreshment of expiration dates. The following questions are taken into account: Do the systems easily allow to classify and separate

archival data from master data items. Which mechanisms are in place to handle data retention and deletion controls and at which degree of granularity. Is it for example possible to connect to external storage systems that ship with built in mechanisms for data retention? Which security mechanisms for supervising deletion control mechanisms are in place?

CHRONOS ships with modules for creating archival data retention policies and fully applies to the requirements of implementing legal hold within a repository. There are mechanisms in place for interacting with database environments themselves but primarily data retention policies are enforced on the exported data. The software allows to enforce retention and deletion policies across different storage media and provides a central interface for maintaining distributed archival packages. CHRONOS offers adapters for interacting with dedicated storage facilities as for example provided by EMC[2]. The system not only takes advantage and closely integrates with these advanced storage technologies but also provides retention mechanisms for standard file volumes which don't offer out of the box capabilities for defining update strategies, expiration dates, etc. The degree of granularity on which the system is able to operate upon is a single archival package. The actual process of marking data for deletion and enforcing the physical deletion of data from the media is a two step process and is safeguarded by human approval with dedicated access rights.

SIARD, by design, exclusively offers support for the database retirement scenario. All information the application is able to access within a database gets archived and it is up to the user to provide adequate visibility and access right policies to the targeted data sets via the database's management component. SIARD never writes or deletes information to or from a database system as it is executed with read-only access permissions. All information is written to the standard file-systems with no SIARD internal support for data retention or different storage connectors. Data integrity as written to the file system is guaranteed by the SIARD-Suite, but it is up to the archivist to take care of everything beyond.

Support of UDTs and Oracle Specifics

Clarifies the degree of support for custom Oracle database features such as user-defined datatypes (UDTs) or Oracle specific extensions as PL/SQL, Oracle Spatial and custom built applications with Oracle Forms.

CHRONOS is able to archive cascading Oracle user-defined datatypes in a preliminary form and CSP has announced further support for upcoming releases together with comparable constructs of other database vendors. However UDTs are seen problematically given their inconsistency and incompatibility across different versions of Oracle databases. UDTs are only available for current Oracle product versions and only when the JDBC driver offers support, no cross vendor mapping is possible when re-importing archived data. To gain performance in Oracle it is possible to temporarily disable the checking of foreign key constraints when importing a large datasets. This state is not reflected in an exported CHRONOS archive and therefore falsely enabled as active foreign key when re-imported. *In the process of data export CHRONOS makes use of native dialects for querying the individual database systems.* CHRONOS itself delivers a SQL92 interface for running queries on archived data. *Procedural Language SQL (PL/SQL) is neither supported for querying nor for archival purposes.* Additional Oracle specific extensions as Oracle Spatial are currently not supported. For form based applications such as created through Oracle Forms CHRONOS is able to act as middleware through its provided APIs.

SIARD supports standard 'scalar' SQL data types (Strings, Numbers, Dates). There are *plans to enhance the SIARD format to accommodate user-defined data types (UDTs) in a SQL99 standardized way*, in order to fulfill backward compatibility requirements of the SIARD format. There are no plans to further support other Oracle-specifics with one exception, the export of Oracle table and column comments as metadata comments. SIARD's product design focuses solely on standardized data content, which in a SIARD understanding is the only amenable way to long-term preservation. Additionally, vendor lock-in of any kind is avoided in this way.

Rights, Roles and User Management

Access controls and user management is a core component of a running database environment. This section focuses on the capabilities of the tested database archiving products to offer rights, roles and user management functionality on top of the extracted database archive.

CHRONOS delivers a mature user, rights and access management layer out of the box. It is tightly integrated throughout all delivered CHRONOS components and is highly customizable to individual needs. Integration of central user management systems like LDAP is possible. The provided level of granularity allows to protect sensitive data in the archive at the level of database columns.

SIARD itself does neither provide user, rights or permission management nor custom application views within the user interface on top of the underlying data but rather makes extensive use of the underlying RDBMS user management component. Visibility and access rights of the archiving user determines the scope of harvested data as SIARD performs a full database export of all 'seen' objects.

Archive Data Access and Performance

One of the core features of CHRONOS is the system's possibility to execute SQL statements directly on top of the archived data located on the file system with performance measures comparable to those of standard database systems. *To overcome the bottleneck of finding, accessing and processing archival packages from the file system CHRONOS makes use of a hybrid approach of a custom SQL92 interpreter, global search index and a local BTree index on column level, as well as H2 and hsqldb in-memory database systems for SQL JOIN operations.* SQL queries without pre-processed indices i.e. a full archive search, are possible but not very performant therefore the data selection for pre-indexation is essential. A core parameter for adjusting performance in CHRONOS is the archive package split size. This allows to decide how to allocate tabular content into different physical

zip containers. Finding an optimum balance between data access and search is highly use case dependent.

With the CHRONOS database archiving product suite it is possible to both create a database export in a vendor independent generic format that from a technology point of view does not contain any crucial dependencies but roughly just data and corresponding schematic structure. At the same time the software suite delivers added value on top of the physical archive which is crucial for the management and use of such information. For example performing queries over different revision of data, i.e. search operations on content at a given structure and point in time which are performed directly upon archival packages on the file system, expressed in SQL92 and in performance that we're used from database operations. And all without having to re-import and revive archived data in a dedicated database environment and even if in the meantime modifications on the database schema have been undertaken.

SiardEdit is a graphical user interface application for exploring SIARD archive files. *SiardEdit is the central instrument with which SIARD formatted data is processed.* It allows to display, sort and browse primary data in a SIARD archive and to add to or change the archival metadata. Primary data cannot be changed. However the tool is not suitable for complex research or research within large archives. In this case it is recommended to load a SIARD archive into a database system and use database techniques for exploration.

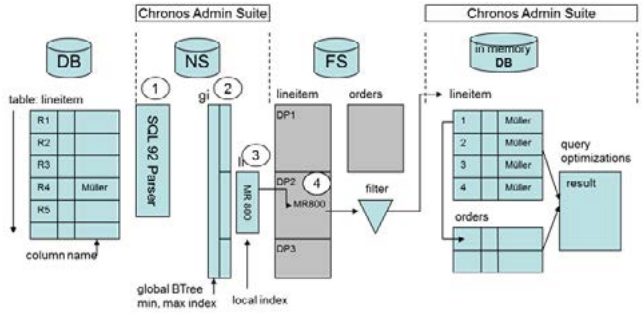


Figure 1: Simplified scenario of a SQL JOIN operation between two tables within CHRONOS depicting the interaction between the custom CHRONOS SQL92 parser, BTree indices for data retrieval, in memory databases for the JOIN operation. Only the indices requires to be on a near storage (NS as disk) next to the search server to execute the query, the actual archival data may be distributed across multiple backends and far storage units (FS as tape).

Syntactic and Semantic Data Changes

In the case of continuous archiving partial datasets remain within the production environment. Therefore a common scenario which needs to be dealt within is the reaction to syntactic and semantic changes over time. Which form of support or traceability do the systems provide for this kind of temporal changes?

Structural changes in the schema as adding additional columns, are automatically detected by CHRONOS. Data is exported into a separate revision and for more complex changes the user is given tools to administrate them. When running a query against a given revision CHRONOS only takes the structure and data into account which was present at that time. *Semantic changes always require manual treatment as there is no way for detection. CHRONOS offers support to automatically transform deposited data via customizable operations for an entire revision.* Those script based operations are written in Java and allow to use the full richness of the JDK for data manipulation. The actual physical content within the long-term archive however stays authentic and untouched as semantic transformations are only reflected within the CHRONOS middleware. The content of a given revision is therefore always properly and consistently reflected on the file system in the state it was extracted from the original database. Duplication of data between revisions is deliberately accepted.

SIARD cannot be evaluated against this use case as it exclusively offers support for the database retirement scenario. The tools is neither designed to cope with semantic or syntactic changes of the underlying data nor does it provide support for handling modifications in the archived packages within SIARD Suite.

Existing APIs and Interfaces

The scenarios archiving, data access and search were evaluated with respect to available programming interfaces.

All of the CHRONOS server modules offer programmatic access via JDBC, Java RMI and web-services and allow deep system interaction. JDBC drivers provide unified access to database systems out of a Java environment. CHRONOS provides direct access to previously archived content on the file system through a JDBC class 4 driver and therefore allows to easily select and process data. Data manipulation is not possible via JDBC. From a functional point of view available interfaces have been tested to programmatically support the entire process of setting up and running a database export and re-importing a CHRONOS archive into a database system. The range of available interfaces differs depending on licensing. Beyond this there is out of the box support for a variety of external facilities such as job schedulers as crontab or taskmanager, storage solutions like EMC^2 centera.

SIARD is a generic platform-independent JAVA program that achieves a lot of independence from the individual database system by being bound to the JDBC interface. As interfaces to SIARD Suite the two command line applications SiardFromDb and SiardToDb are provided for extracting a database archive within the SIARD format or vice-versa. Although the applications' functionality is identical with the functions available via SiardEdit it is recommended using the command line versions especially when downloading large databases as they are designed for scalability. All settings for those tools can be provided via a configuration file, so using the two applications within scripting solutions allows to achieve a certain degree of automation as e.g. scheduling via cron jobs. All surrounding dependencies need to be configured externally.

Scalability and Limitations

This point takes into account scalability aspects as size, throughput, access time as well as any form of limitations that could influence the products usage as hardware and software prerequisites.

CHRONOS is a product which in all aspects is designed to deliver performance and scalability via Java multithreading. In our testing environment with standard hardware running 4 CPUs and 6 GB of RAM we were able to constantly export two thousand tuples per seconds from the database even running the archival packages indexing operations aside. The bottle neck in this case was the performance of the underlying local file system. Scenarios with limited memory resource allocation of the Search- and LocalIndexJobs can noticeable bring down the response time of the system whereat only 128 MB of assigned Java heap memory still were sufficient to properly execute operations on the SQL search server without any erratic behavior.

Both CHRONOS and SIARD are self documented archives of primary data. External documentation, artifacts, process documentation, approval or decisions taken are not part of a created archival package even though this information is partially available through the software suite. Due to integrity checks of the archival zip packages it is not possible to add this information externally.

The ZIP 64 standard accommodates files with sizes up to 18'446'744'073'709'551'616 Bytes (i.e. 16 Exabyte). *SIARD uses ZIP 64 without compression to generate a one-file container for the archived database and is therefore limited by this size.* The SIARD Suite runs within a JVM of typically 500-2000 MB of heap space. It uses the heap space for holding all metadata in memory as well as one row of data. This JVM setup has been sufficient for any database tested. SIARD does not make use of JAVA multithreading or multiple DB sessions due to the imposed number of integrity problems! While the Java Swing application SiardEdit had memory problems when downloading a large number of items, the provided command line applications showed consistent performance.

Risk Behavior and Dependencies

Whats the degree of underlying dependencies for a given database archive in subject to system dependencies, vendor / tool locking, or similar objectives?

Both tools follow the approach of clearly separating the composition and description of the data structure from the actual primary data - this is also reflected on file system level. CHRONOS describes the structure in XML and provides a fully interpretable XSD schema file while the content itself is stored in a delimiter file. In theory all information to properly read and interpret this data and therefore possibly manually revive it into a RDBMS in case of a vendor crash is available without any direct dependencies. In practice this step is non trivial and not possible out of the box without a previous data transformation process due to the fact that both SIARD - which makes this fact implicit by proposing a central data exchange format and representation based on SQL99 - but also CHRONOS require a mapping between their internal form of data representation and the corresponding database datatype configuration and mapping. While within SIARD this commitment and scope is based on SQL99 datatypes to guarantee a full round-trip scenario, CHRONOS explicitly documents the supported datatypes for every vendor and database version but however treats the cross-vendor and inter-version representation as industrial secret.

A main aspect in digital preservation is to keep the stack of software dependencies as low as possible. For CHRONOS they can be mainly summarized as Java + JVM, XML and Zip32 Deflate. The zip compression deflate is public domain and widely used as for example within the Portable Network Graphics (PNG) or ISO Open Office XML-Format. Additional system configurations, documentation regarding the technical approval processes as the underlying user, role and rights management are not part of an archival package but partially are reflected in the applications settings in XML form. It should be possible to enable manual database recovery within a fair amount of time.

Referential Dependencies

In many cases the database does not contain full referential integrity as this is often depicted by external documentation or reflected within a different software layer. In some use cases it many be required to export a given dataset including all referential dependencies? *CHRONOS allows to automatically detect referential dependencies for master tables and has tools to decide how to deal with cyclic references and to with depth those references need to be respected. External dependencies can be remodeled.*

SIARD has the ability to archive an entire database, but without the possibility of selecting individual tables. However the 'entire database' refers to the collection of all objects that the database user which is used to export the archive has read access to. Therefore to exclude certain tables from the archival process the only additional step required is to create a database user with specific read access rights limited to the tables that should be archived. All foreign keys are resolved if the SIARD file was generated from a database which had constraints enabled. SIARD does not censor data or ensure integrity.

Standards and Compliance

Currently there is no standard in the field of long-term archiving for databases. The SIARD format has become a widely accepted format for the exchange of relational database content within GLAMs.

Is there a chance for an SQL standard for Archiving, based on a subset of the ISO-9075-SQL, similar to the PDF/A for archiving? To increase acceptance by vendors the SQL standard defines three levels of conformance and implementation: entry, intermediate and full level. The mandatory part of SQL99 is called core and is described in part 2 (foundation) and part 11 (schemata) of the standard. Since most RDBMS are based on SQL and most vendors claim compliance with the standard one should assume that relational database definitions are independent of any specific RDBMS product. Unfortunately this is far from the truth. Even though the SQL standard today comprises over 2000 pages it is far from being fully self-contained. In contrast,

SQL99 explicitly identifies 381 so called implementation-defined items. Most of today's RDBMS implement (and sometimes faultily) only the core and the entry level of the standard completely. To this often large number of non-standard, product-specific enhancements are added which leads to many different SQL flavors. *SIARD Suite currently adheres to SQL:1999 "Core Features" in terms of supported functionality and mapping of data types.* Future versions may be extended to make use of additional SQL99 components as Packages.

The OAIS model is a reference model for a repository where a SIARD archive would be a Digital Object held within an OAIS repository. The SIARD archive therefore is not a stand-alone single file that can be though of as an AIP. A SIARD file should be treated as a single object – like a word file – in an archival system, which itself may or may not adhere to the OAIS model. It is assumed that a retired database in the SIARD format is archived as part of a larger archive package with additional documentation. In the case of SIARD, it is important to separate the discussion of the format from the discussion of the tool. The formats huge advantage is that it is solely based on existing international standards and independent from any single database vendor or the specific infrastructure of a particular customer. The SIARD tool has more disadvantages. It makes assumptions and decisions about the mapping of real live databases to the standard. These may be questioned. However, this is not a failure specific to the SIARD software. The author is not aware of any tool that explicitly guarantees the preservation of primary data values and idempotent up- down and -uploading. The tool creators have made the decision to prefer moderate performance over database or operating system dependence. The existence of this "reference implementation" does not prevent the implementation of other solutions with higher performance or even with vendor lock-in.

Structure, Setup and Size of the physical Archive

The Transaction Processing Performance council database dump was used to get measures and comparison on the physical size of an exported database archive. Not taken into account in this comparison are parameters which are built up within a database environment that are not easily uniquely assignable. The size of the original source of a database is not a defined value i.e. there is no measurement on the size of an Oracle schema or database index in bytes?

While a SIARD archive required +338% on disc space compared to the database dump a CHRONOS archive is able to decrease the required space by -41%. This comparison took into account operational artifacts which are understood and processable by SQL-Developer, SIARD-Suite and CHRONOS Administration Suite. SIARD uses a zip container but does not apply any compression algorithm. By applying a post-compression (deflate, 32K word size, standard compression) the size of a SIARD archive can be brought down by 30%.

For CHRONOS in average we measured a 40-60% reduction in required file size compared to the database dump depending on the underlying tabular data. Further room for improvement lays in the use of different checksum algorithms. MD5 is applied out of the box and tends to blow up small records. As the 32-Bit version of a zip container is only able to support container file sizes up to 2 GB the system splits up archival packages. Per default 20 MB is the standard package split size which also shows the best performance stats regarding searchability, indexing and query response time. The file structure of an exported database archive within CHRONOS separates the actual tabular data from its structural description. Partial retirement scenarios are built up based on temporal events, either static or based on temporal markers within the database. Elements as Binary Large Objects (BLOBs) or CLOBs are stored in separate clusters of binary objects within the archive and are referred to via data pointers in the tabular data. Data integrity against the original is checked by the system after harvesting as well as after moving the archival data into the storage component.

For primary data SIARD chooses to use XML short tags. In our TPC-C test data we were able to notice a factor of 1:3 of increase in data size. According to SIARD's official statements the size of the SIARD file should be similar to the size of the Oracle dump from which it was downloaded, if the primary data represents the majority of information. Even though zip64 packaging is used to create the container file, no compression algorithm is applied to avoid any dependencies for the long-term. The SIARD format is not configurable in the sense of being able to add additional fields. The Swiss Federal Archive feels that an international standard is better served by uniformity than by flexibility. The 'technical metadata' describing the database structure is dictated by the SQL standard. Once a SIARD archive is exported its consistency with the underlying database's data is verified and the number of archived records is documented. Any modification to the database during the process of creating the SIARD export leads to an error in the exporting process. Regarding SIARD's structure on the file system, all database contents such as schema definitions and primary data are stored in a collection of XML files which conform to the SQL99 definition. The only exceptions are binary large object (BLOB) and character large object (CLOB) elements which allow holding larger sets of data. These are stored in separate binary files having referential pointers in the corresponding XML entries. Data is stored in a Unicode character set. While extracting databases that support different character sets, a mapping to the corresponding Unicode characters is carried out. For this reason, SIARD generally translates national character string types in the database software (NCHAR, NVARCHAR and NCLOB) into non-national types (CHAR, VARCHAR and/or CLOB). This convention is well supported by XML and independent of whether an XML file is stored in the UTF-8 or UTF-16 representation. Characters with a special meaning to XML are substituted by entity references in the SIARD archive files. If a string is longer than 4000 characters then „clobType" and „xs:string" are replaced by an external reference to a text file. If a binary array is longer than 2000 bytes then „blobType" and „xs:hexBinary" are replaced by an external reference to a binary file. Characters that cannot be represented in UNICODE as well as the 'escape character' and multiple space characters are escaped as ŏ0<xx> in the corresponding XML, 'greater than', 'less than' and ampersand characters are represented as entity references in XML.

Specification of Information Lost

Which audit trail capabilities does the system offer for logging and tracking modifications over time. Is there a way of specifying the amount of information lost when exporting data into a long-term archive? One example on a measure which could be applied is the Oracle SQL Minus operation after re-importing a database archive to determine the correct structure and item count against the original data.

The amount of information available in the database's metadata is debatable and cannot be quantified. Both SIARD and CHRONOS can be classified as idempotent in terms of that an upload – download – upload produce delivers the exact same data types and values on the second upload as on the first one. Checking this idempotence is part of the SIARD build script. However there is no support for statements that declare what information is actually lost during export (as e.g. UDTs, disabled Oracle foreign key constraints, etc.), lost during cross database or cross db-version re-import or lost by a mapping from the native type to SQL99. Both CHRONOS and SIARD support program logging with various log levels to track down system behavior but no persistent logging of the history of changes is implemented. SIARD per definition does not support schema changes or ongoing database archiving over time and takes an archived/retired database as a final and unmodifiable constructs there is no need for data modification audit trails or similar tracking features at this level. Features like these are more in the realm of the enclosing archival process/system and therefore a feature which one would expect in system like CHRONOS.

4. CONCLUSION

Archiving databases either means preserving information or preserving functionality or both, so the tools SIARD and CHRONOS were evaluated within the scenarios of database retirement, continuous/ongoing and partial retirement as well as application retirement. In the underlying case study both tools proved stable and technically mature in creating a database archive in a vendor independent long-term preservation format for a rich number of relational database system. The tools proved mature and were able to deliver solid performance. There are small differences in the number of supported database vendors, SQL elements and the internal data representation. A clear recommendation which product the community should adopt is almost impossible as the supported scope and use cases both tools are able to deliver are highly diverse. SIARD suite was designed as reference implementation for the SIARD format and exclusively offers tool support for the use case of database retirement. CHRONOS on the other hand, a commercial product well designed for scalability and industrial needs, provides a rich set of tools and end-user applications that allow both to export a physical database archive and to operate on top. CHRONOS provides all mandatory bits at the required level of complexity to accomplish the challenges of the ongoing/continuous and partial archiving scenario. Core features include running SQL92 queries on top of the archived data with database like performance, support for revisions, syntactical and semantical schema modifications, resolving cyclic dependency, external referential integrity handling, a full blown access control and data retention layer, etc. Shortcomings of CHRONOS are the limited support of complex objects as Oracle UDTs and lacking support of audit trails for classification and documentation of information lost. Besides the core functionality CHRONOS provides support for the use case of application retirement with tools that allow re-modeling of business objects, application logic and reporting functionality and by being able to directly serve as middleware layer for legacy applications. The rich set of programmatic interfaces allows both to integrate with most of the system's functionality as well as to grant access to data via standard mechanism as JDBC. Finally we presented examples why preserving complex structures as databases through a record centric approach does not only solely depend on the amount of information captured from a database itself but why it is important to create full preservation packages which cover contextual information.

Acknowledgments

The author would like to thank Mario Günther, Mario Täubler (CSP GmbH Co. KG) and Thomas Hartwig (Enter AG) for their input on the underlying study. Their comments, written feedback and interviews helped to clarify open issues in the process of evaluating CHRONOS and SIARD. Any remaining misinterpretations or mistakes are those of the author.

5. REFERENCES

[1] Agrawal, R., et al.: The claremont report on database research. SIGMOD Rec. **37**(3) (September 2008) 9–19

[2] Edelstein, O., Factor, M., King, R., Risse, T., Salant, E., Taylor, P.: Evolving domains, problems and soluations for long term digital preservation. iPres (2011)

[3] Schmidt, R.: An architectural overview of the scape preservation platform. iPres (2012)

[4] Heuscher, S., Stephan, J., Peter, K.M., Frank, M.: Providing authentic long-term archive access to complex relational data. CoRR (2004) DL/0408054

[5] Brandl, S., Keller-Marxer, P.: Long-term archiving of relational databases with chronos. First International Workshop on Database Preservation (March 2007)

[6] van Essen, M., de Rooij, M., Roberts, B., van den Dobbelsteen, M.: Database preservation case study: Review. IST-2006-033789 Planets Deliverable PA/6-D13 (2011)

[7] Brown, A., Lappin, J.: Ecm talk 17: Practical digital preservation (2013) `http://traffic.libsyn.com/ecmtalk/ECM_Talk_017.mp3`.

[8] Burda, D., Teuteberg, F.: Sustaining accessibility of information through digital preservation: A literature review. Journal of Information Science (2013) 1–19

[9] Becker, C., Rauber, A.: Decision criteria in digital preservation: What to measure and how. Journal of the American Society for Information Science and Technology (JASIST) (2011)

[10] Strodl, S., Petrov, P., Rauber, A.: Research on digital preservation within projects co-funded by the european union in the ict programme (2011)

[11] Farquhar, A., Hockx-Yu, H.: Planets: Integrated services for digital preservation. International Journal of Digital Curation **2**(2) (2007) 88–99

[12] Rammalho, J.C., Ferreira, M., Faria, L., Castro, R.: Relational database preservation through xml modelling. Extreme Markup Languages (2007)
[13] Stefanova, S., Risch, T.: Scalable long-term preservation of relational data through sparql queries. Semantic Web Journal
[14] The Danish State Archives: Symposium about the transfer, preservation of and access to digital records based on the danish experiences (2008)
[15] Ribeiro, C., David, G.: Database preservation briefing paper
[16] von Suchodoletz, D., Rechert, K.: Migrating of complex original environments - verification and quality assurance challenges. JCDL (2013)

File-Based Preservation of the BBC's Videotape Archive

Thomas Heritage
BBC Research & Development (R&D)
Centre House, London
W12 7SB, UK
thomas.heritage@bbc.co.uk

ABSTRACT
The BBC Archive now contains around 15 Petabytes (single copy) of uncompressed audio-visual files that have been created from videotapes since 2007. This process is still on-going, creating an ever growing file-based collection of the BBC's television history. This is of course in addition to the new content now being produced that begins life as files. This paper focuses on the technology aspects of the digital preservation of the file-based historical TV collection and looks at how this currently isolated collection may later interface with other systems and collections. Consideration is given to what has been achieved so far, some lessons learnt, and the future challenges.

Keywords
Television, Digitisation, Preservation, Migration, Archive, Library, MXF, OAIS, LTO

1. INTRODUCTION
The BBC Archive contains more than 12 million items including several million television items held on either film or videotape [1]. Migrating the content from physical carriers to files ensures the preservation of content previously held on obsolete carriers, reduces the physical storage space required for the collection, and brings about new opportunities for providing access to the archive. In 2007 the BBC began creating master media files from television content held on Panasonic D3 videotapes (a process known as 'ingesting') [1].

The systems have since been developed and are now in use at the BBC Archive Centre in Perivale (West London) for the ingest of Sony Digital Betacam (DigiBeta) videotapes. So far, around 100000 D3 and 125000 DigiBeta videotapes have been ingested representing about 15 Petabytes of content (single copy). It is these videotapes that are considered here. The processes that have been / might be applied to television content held on other videotapes, film, etc are not considered in this paper (e.g. the collection of U-Matic tapes that were migrated to MPEG-2 files stored on DVDs).

The digital preservation of this content is carried out by the BBC Information & Archives department, with many of the systems and processes developed in collaboration with BBC Research & Development and other partners. An overview of the current status is given in Section 2 with a description of the core processing systems, consideration of the 'level' of digital preservation that has been achieved, and some of the challenges faced and lessons learnt. With a large proportion of the content held on LTO3 data tape, action will soon be required to migrate this to a new storage technology before these tapes become difficult to read – the issues involved are considered in Section 3. The focus so far has principally been on preserving, as files, the content that was held on videotape. However, with more production facilities operating completely tapelessly, providing file-based access to the preserved content is an important area to address. With this in mind, the content migration from LTO3 data tape will need to be considered with regards to the wider context of file-based archives and production systems. A simple model is presented in Section 4 of how such systems are likely, in practice, to relate to the long-term preservation collection of historical TV content.

2. THE CURRENT STATUS
2.1 Core Processing Systems
Figure 1 provides an overview of the core systems involved in the ingest of videotape content to files and the subsequent preservation operations (databases, reporting systems etc are omitted). Further details of each of the systems are given below.

2.1.1 Preparation (DigiBeta Only)
Prior to ingest the videotapes are checked and prepared physically and checks are made on their metadata. Videotapes containing *content* that the system deems should not be ingested are rejected: this may be because the content has already been ingested successfully (from this, or another, videotape) or has not been selected for preservation. Videotapes to be ingested are rewound and any paperwork, barcodes, etc with the videotapes are corroborated with each other and with the system. Any videotapes requiring metadata correction or enhancement are removed to be dealt with separately. Those videotapes that remain are sorted by the number of content items that they hold and by video aspect ratio, ready for ingest.

This level of preparation ensures that the ingest process is as smooth as possible such that an efficient 'preservation factory' [2] is established.

2.1.2 Ingest
The ingest function is performed by the Ingex Archive system originally developed by BBC R&D [1][3]. For the transfer of D3 videotapes custom hardware was designed to convert the data on the tapes to a practical form at the highest quality [4]. For the transfer of DigiBeta videotapes custom software has been added to verify the performance of the Video Tape Recorders (VTRs) as well as to detect video faults specific to the DigiBeta format.

After a videotape has been recorded, 'chunking' is performed to split the data such that one master media file is produced for each content item. A brief review of each file then takes place, principally based on any errors or features that have been automatically detected during ingest. When enough master

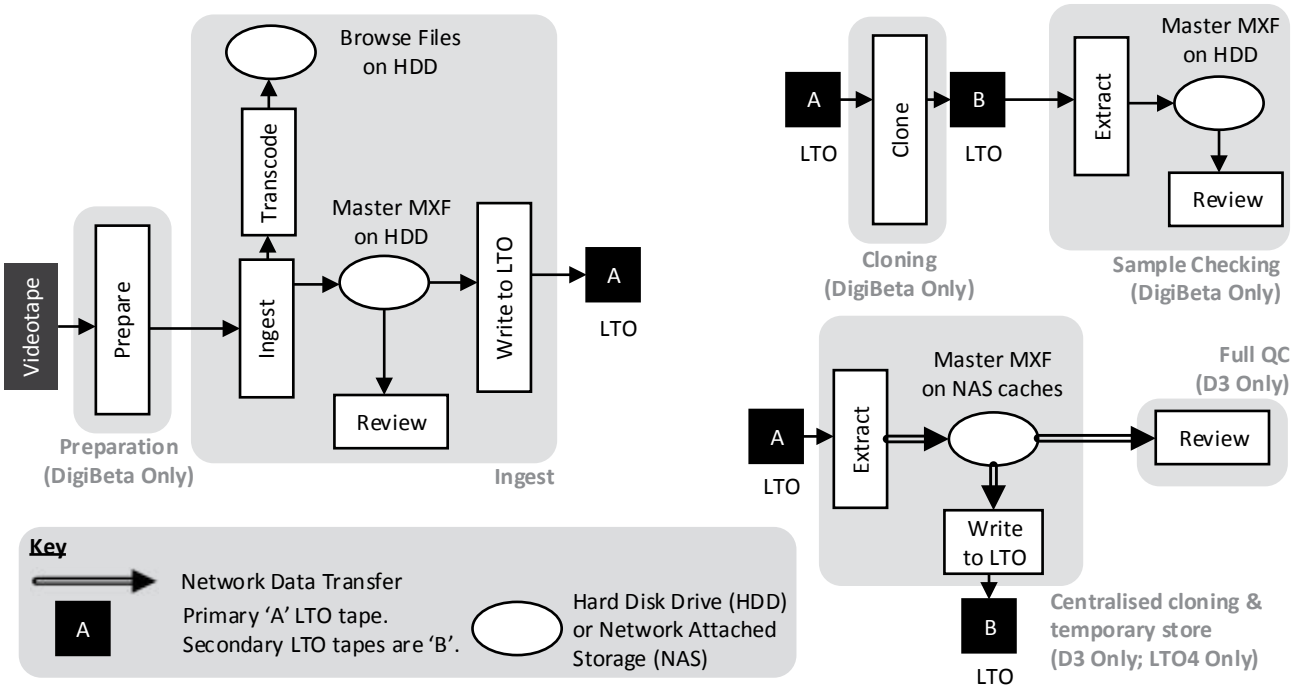


Figure 1. A simplified view of the core processing systems for D3 and DigiBeta videotape content preservation.

media files have accumulated they are written to LTO data tape (LTO3 originally; LTO4 currently). The MPEG-2 browse files are produced locally during ingest and later 'harvested' to a central location. They have been used to provide some access to the collection as well as to aid in Quality Control (QC).

2.1.3 Cloning (DigiBeta Only)

This LTO cloning system operates by cloning files directly from input LTO tape to output LTO tape. This avoids any need for a storage cache inside the cloning machines which simplifies the system, reduces the cost, and removes another source of possible errors and system failures. The supported LTO generations are governed purely by what the tape drives support, and the capacity of the system can be scaled simply by adding additional cloning machines. Although consideration was given to modifying the ingest system to produce an additional copy of each LTO tape, the present arrangement was chosen as it ensures that every 'A' LTO tape is fully read and independently verified.

2.1.4 Centralised Cloning & Temporary Store (D3 Only; LTO4 Only)

This system provides:

- *LTO tape extraction & writing.* Content from LTO tapes is extracted onto Network Attached Storage (NAS) hard drive caches and new LTO tapes are produced from this content. LTO tapes are out of the vault for the minimal amount of time and are only handled by the logistics staff.
- *File playback over network.* The master media files extracted from LTO tape can be played over the network from any machine allowing QC of the master media files by both the primary operators and their supervisors (who previously had to rely on the browse files).
- *Automated management.* The system itself and the processes it supports are automatically managed thereby simplifying workflows. QC operators are automatically allocated content to be reviewed rather than allowing them to choose – if the operator chooses content they are interested in then less attention is paid to the technical quality.

The centralised nature of this system naturally provides some disadvantages also. Principally, the workflows of the associated operations are less flexible, and the system component that manages its operation constitutes a single point of failure. Additionally, the need for NAS caches introduces the issues avoided for LTO cloning as described in Section 2.1.3.

2.1.5 Sample Checking / Full QC

A sample of DigiBeta content is checked by extracting from LTO tapes to a hard drive cache on the review machines. All D3 content is fully manually QCed either by playback over the network or by extracting from LTO tape onto the review machine as for DigiBeta content (latter case not shown in Figure 1). The review processes provide feedback on the integrity of the LTO tapes, metadata accuracy, master file technical details, errors introduced to the content by the ingest process, etc.

2.2 Archival Information Packages (AIPs)

It is instructive to assess the preservation outputs in relation to the OAIS concept of an Archival Information Package (AIP) where the LTO tapes constitute the Archival Storage [5] (only the newest format preservation outputs are considered). Firstly, it is useful to consider the elements that *delimit* and *describe* the AIPs:

- *Packaging Information.* The master media files are Material eXchange Format (MXF) OP1a adhering to a custom BBC Archive profile [1]. They are written to LTO tape (without compression) following a custom scheme using TAR archive files and plain-text index files. So, each

AIP is actually a combination of elements from the LTO scheme and MXF profile. Each AIP is identified by the MXF filename (held inside the TAR archive, the MXF file itself, and the LTO index file) and the MXF Unique Material IDs (UMIDs).

- *Descriptive Information.* This consists of: metadata from the AIP (such as Programme Title); content properties such as duration; browse files.

The AIPs themselves consist of:

- *Digital Object.* It is valuable to realise that the Content Data Object to be preserved consists of the bitstreams representing the audio, video, and timecode from the source videotape rather than the entire MXF file (which additionally contains many of the metadata items identified below). The audio and video are stored uncompressed, immediately achieving the migration end-state promoted by PrestoPRIME [6].
- *Representation Information.* The LTO scheme is described in the plain text index files on the tapes themselves. The MXF profile is fully described in PDF documents [1] – these depend on numerous other documents (e.g. MXF standards) and are not stored in the AIPs.
- *Reference Information.* Includes the programme title etc and content identifiers such as the BBC 'programme number'.
- *Provenance & Context Information.* Details are included of: the original content transmission date etc; the videotape the file was produced from; the ingest process.
- *Fixity Information.* Checksums of the MXF files and the LTO index files are stored on the LTO tape. The MXF files contain checksums per frame for each audio / video track.
- *Access Rights Information.* Any details that may be available are stored in completely separate systems.

2.3 Digital Preservation Assessment

The systems and processes setup to preserve content from videotape did not set out to establish a complete Trustworthy Digital Repository [7] – instead the BBC decided to "get on with it" [8] and focus on transferring content from videotape while the machines were still available (especially a concern for D3). Over time the systems have evolved and been added to in order to support additional videotape formats as well as to introduce improved digital preservation practices, and they will continue to evolve in the future. The concepts of "Levels of Digital Preservation" [9] and "Digital Archiving Maturity" [10] are quite useful in understanding how digital preservation systems can evolve through stages.

Table 1 shows *highlights* of digital preservation developments and gaps in relation to the BBC's historical TV archive. It is certainly not comprehensive (e.g. as in [7]), completely ignoring issues of funding, administration, preservation planning, etc (these are ignored not least because aspects of these elements are common to other areas of the BBC Archive including to collections that are not file-based). The main changes have been due to the introduction of fixity information and improved documentation (Section 2.2), the introduction of cloning (Sections 2.1.3 & 2.1.4), and additional work on database integrity and data reporting (Section 2.4.3). Possible future developments are discussed in Sections 3 and 4.

Table 1. Digital preservation development over time.

	2007	**2013**
Full AIPs	+	++++
Number of AIP copies	1	2
Regular object fixity checks	N/A	No
Provide access to content for re-use – on videotape	Yes (manually)	Yes (manually)
Provide access to content for re-use – as a file	No	No
Data Management	+	++

'+' indicates advancement towards an OAIS.

2.4 Main Challenges & Lessons Learnt

2.4.1 Custom Designed Systems

All the systems described in Section 2.1 were custom designed with custom schemes for writing the master media files and LTO tapes as described in Section 2.2. Using custom solutions has allowed the BBC Archive's precise requirements to be met which was not felt to be possible (at least in 2006 / 07) using solutions available in the industry. It has also meant that there is: complete transparency as to how the systems operate; no reliance on a third party solution provider; the option to add new features as required. However, a firm commitment is required to support and maintain the software (even today issues are being discovered with software first developed six years ago) and to produce tools (and new systems) to operate on the custom MXF / LTO schemes because they are not fully supported by industry solutions. The work required in testing and documenting the systems should also not be overlooked.

2.4.2 File Sizes, Data Rates & Storage

The main limiting factor in many of the systems is data input / output (and consequently data movement times) due to the master media files (MXF) being 75–100GB per hour of content and the fact that around 30TB of content can be produced per day by the current DigiBeta ingest process at Perivale (24 stations running simultaneously) – this is clearly one of the disadvantages of storing uncompressed content. Such data rates require a different approach compared with systems handling small files. For example, the only practical solution to producing checksums for the MXF files is to construct a processing pipeline that does this while simultaneously performing other operations (such as copying files from hard drive to LTO tape). Moving or playing-back these large files over a network (Section 2.1.4) requires careful consideration of the design of the network as well as the whole software stack on both the NAS and the access client. For example, the MXF player software built by BBC R&D and used for file review over the network had to be modified in order to prevent overzealous file caching that could use up all available network capacity. Inside the ingest system (Section 2.1.2) the data rates are also a major challenge. For example, 'chunking' involves reading back the recorded content while writing new files (one per content item). If LTO writing of other files is happening simultaneously then

the system is obliged to severely limit the rate of chunking which introduces a large delay before the next videotape can be ingested. This situation could be dramatically improved by ingesting content items separately, an approach that is now enabled by the collection of item timecodes prior to ingest (an example of 'metadata enhancement' described in Section 2.1.1).

All the LTO tapes produced by the preservation systems are handled manually and stored on shelves / in crates with 'A' and 'B' tapes stored in different locations. This is a flexible solution (compared to storing tapes in robots) that fits well with the skills and facilities already present in the Archive for handling other physical assets such as videotapes. It also means that all the content is completely offline and can be moved at high speed around the Archive (consider the 'bandwidth' of a trolley full of LTO tapes!). However, it does mean that: there is no automatic management of tape locations; access requires human handling which introduces delays and exposes tapes to less than ideal conditions; tapes have to be treated as 'units' rather than being able to handle the contained files individually.

One observation of MXF file corruption highlights that while checksums are important it is critical to understand at what point in the file's life they were produced. In this case, a number of MXF files stored on LTO tape were found to be corrupted towards the end-of-file due to an issue with the hard drives in the ingest station while the files were being written to LTO tape. Given that the checksums are produced during the tape writing process the actual checksums of these corrupt files (as stored on LTO tape) matched the expected values. Alterations have since been made to detect such hard drive errors. An even more robust solution would involve verifying the internals of the MXF files as they are written to LTO tape including the per-frame checksums.

2.4.3 Metadata & Databases

No databases are shown in Figure 1 but they are of course a crucial element. Some of the main challenges & lessons include:

- *Many databases.* Metadata is distributed between numerous databases, most of them with different schemas. However, this is still preferable to data being stored in text files etc as long as a database with standard query interfaces is used.
- *Duplicated data.* Sometimes this is helpful but it can make correcting any metadata errors very difficult to do correctly.
- *Missing fields.* New metadata fields to aid in error diagnosis are continually being thought of but in many systems it is not straightforward to add them.
- *Data integrity.* As much as possible, integrity should be enforced by the database itself to avoid erroneous data.
- *Data access.* Mechanisms should be built-in as early as possible for reporting and summarising data for users.

2.4.4 Workflows & Processes

The digital preservation "three-legged stool" [11] is a useful reminder that successful preservation is not all about the technology. Some of the largest challenges have been related to the scale and complexity of the physical logistics operation and accommodating 24-hour working (at times) in order to ingest content at the required rate. The ways in which processes have evolved has certainly altered the preservation of content and tracking / recording these process changes, as well as trying to ensure consistency at such a scale, are real challenges. Work was conducted in the PrestoPRIME project to produce a simulation of the D3 preservation process [12]. Building even the simplified model took a number of months highlighting the complexity of the workflow when all factors are taken into account – even then, setting the model parameters realistically is challenging.

2.4.5 A Complicated & Varied Collection

The videotape collection being ingested contains content from as early as the 1930s (which originated on film), and content from the 1950s and later which may have originated on 2" tape and then been migrated to 1" tape then D3 and / or DigiBeta tape, perhaps with multiple videotape copies being made (which are unlikely to be 100% identical). An item of content may be ingested to file multiple times from the same or different videotapes, each of which may have a different provenance and different faults. This results in a very complex collection to manage and it is not always straightforward to derive the best possible copy of each content item from the ingests performed.

3. LTO3 MIGRATION

With a large proportion of the content held on LTO3 data tape action will soon be required to migrate this to a new storage technology before these tapes become difficult to read (due to lack of support by the latest drives). This migration process presents numerous challenges as well as opportunities to improve the preservation of this portion of the collection.

3.1 The LTO3 Tape Collection

Some key statistics for the collection:

- 14000 LTO3 tapes
- 5PB of data
- ~7.5 years of A/V content if played end-to-end
- Only one copy of each file MXF file is stored
- No Fixity Information

3.2 Designing a Migration Process

A custom solution is almost certainly required for the reasons discussed in Section 2.4.1. Some of the processes that it could potentially include are:

- *Validation of existing AIPs and Packaging Information.* The LTO scheme and MXF files could be validated against the relevant specifications, although without fixity information errors in the Digital Object itself would probably not be detected.
- *Migration of AIPs and Packaging Information.* New AIPs could be generated from the old with augmented content e.g. Fixity Information and additional Representation Information could be added.
- *Creation of additional Descriptive Information.* Some metadata items are held only inside the MXF files so it may be of benefit to extract this information and store more accessibly. Content analysis could be performed: even simple analysis could be very useful e.g. determining how many black frames of video each content item contains.
- *Creation of Dissemination Information Packages (DIPs).* While all of this content is being read from LTO tape it would be possible to create a complete collection of DIPs for ingest into another system. This would probably involve

transcoding the uncompressed content to a format more appropriate for re-use.

- *File audit, retention review, repair.* This may be an appropriate time to filter the collection to remove some of the complexities described in Section 2.4.5. However, this would be complex and perhaps involve some risk as it might involve discarding some MXF files.

3.3 Choice of Archival Storage

The type(s) of storage to use and the number of copies of each AIP to store will be affected by the issues discussed in Section 4. However, it seems likely that at least one copy will be stored on LTO tape, perhaps LTO6. Although LTO3 tape will become an 'obsolete' technology, given the popularity of LTO tape, drives will surely be available to read them for many years to come (although they may be scarce, expensive, and difficult to connect to). Therefore, it is worth considering whether the LTO3 tapes should be kept even after the migration: they would serve as an additional copy of the content for use only in extreme circumstances or if a fault with the migration process is later discovered.

3.4 Choice of AIP & Packaging Information

If it is possible to use standards (ideally commonly adopted and well-supported open standards) when defining the outputs of the migration process then this may mean that a custom solution is not required for the next migration process (although if open standards are adopted then it is not precluded). This is in addition to the benefits of increased interoperability with industry tools, other repositories, etc. Even if it is not possible to use such standards for all 'layers', the higher up the 'stack' that standards are used the more that should be possible with industry solutions. For the same reason there may be a benefit to clearly discrete 'layers' unlike the current situation where there is an overlap between elements of the AIP and the Packaging Information (Section 2.2). The principle 'layers' in the 'stack' are listed below along with possible standards (some not yet completed; note that some standards cover multiple layers) to consider – these are listed purely as examples rather than recommendations. The Presto4U project includes an element of work analysing and promoting standardisation for audiovisual digital preservation [13] and should be consulted for more information on this topic.

- *Archival Storage* e.g. LTO tape
- *Packaging Information* e.g. LTFS, AXF
- *AIP 'wrapper' or (virtual) 'container'* e.g. BagIt, METS, MPEG-A PA-AF, AXF
- *Metadata* e.g. PREMIS [14], METS, MPEG-A MP-AF [15]
- *Digital Object* e.g. AS-07 [16]

Those standards / formats not referenced above are described in [2] or [17].

3.5 On-going Preservation

Consideration also needs to be given to how the migrated content will be managed as a collection / repository and issues of regular fixity checking, data management, repository interfaces, access, etc.

4. THE WIDER CONTEXT

To answer all the questions raised in Section 3 requires an understanding of how this historical TV archive could relate to other archive and production systems in the BBC, and therefore how the content might be accessed. Figure 2 illustrates a possible practical model that appears to be developing for at least the short to medium term. In this very simple model only the historic TV digital preservation archive and a digital archive library are included: the latter represents systems used to manage production quality content (both historic and new) and perform day-to-day functions on it such as search and access for re-use, as well as interfacing to all the other systems required to import and export content (e.g. for television broadcast). This represents the shift of the archive to the heart of the content production and delivery processes.

The formats used in the archive library are those suitable for current processes and much of the content will not be in its final state (with new versions being created, and some content perhaps being reviewed and deleted after short periods e.g. one year). Conversely, the content held in the preservation archive will (principally) be in a final state, stored using an archival format (e.g. uncompressed) and selected for long-term retention. The vast majority of Descriptive Information could be held in the library. This arrangement allows the specific and rapidly changing needs of users to be met by the archive library while the preservation archive focuses on the long-term preservation of content. The preservation archive is able to be managed separately to ensure content security, and need only expose a simple (and fairly static) interface that is used by the library. Not least, the separation between the two systems means that the challenge of providing all the required digital archive functions is divided into smaller, more manageable 'modules'.

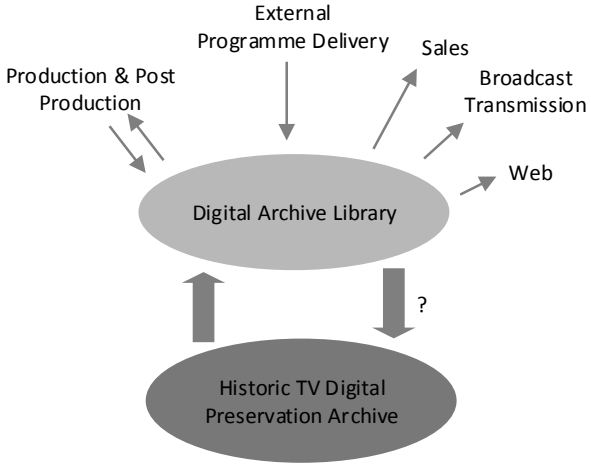


Figure 2. Possible model of a Digital Preservation Archive & a Digital Archive Library

Content in the preservation archive could be accessed on-demand by the library (with the content transcoded to any supported format). However, it may be preferable for DIPs to be delivered in bulk from the preservation archive to the library so that large amounts of content are immediately available to users (but without the same freedom of choice about the format). If content is later required in a different format then a new request can be submitted to the preservation archive. This approach (rather than transcoding from the format held in the library)

avoids concatenated transcodes and so maximises the quality of the content over its lifetime.

It will be important to define the Service Level Agreement (SLA) for the preservation archive [18] as this will help to manage expectations as well as inform design decisions such as the type(s) of archival storage, the repository interfaces, network connectivity, repository management workflows and functions, etc. For example, if on-demand access requests to the preservation archive are to be handled rapidly then it could perhaps be a good option to store at least one copy of the content on (idle) hard drive arrays rather than LTO tape only.

The complete television archive system will be much more complicated than the simplified model presented here. In reality a number of library and preservation systems are likely to exist, each with different functions. These will be easier to manage (and potentially to federate to form a more cohesive archive) if common standards are adopted both for the elements highlighted in Section 3.4 and crucially for repository interfaces and unique identifiers. A central registry of all content may be a key enabler.

A key question to address will be how production quality (compressed) content will be handled e.g. content born-as-files or content ingested from videotape to a non-archival compressed format only. Will this content be delivered straight to the library (as indicated in Figure 2)? Presumably the library would be responsible for the preservation of this content in the short term but in the longer term would it be migrated to the preservation archive (perhaps migrating to an archival format e.g. uncompressed)? Such a decision would probably be required at the library's end-of-life, if not before then. Some preservation strategies and format migrations are explored in [6].

5. CONCLUSIONS

Progress has been made in creating a large quantity of master media files from videotapes with the 'level' of digital preservation developing as the systems and processes have evolved. There is still much work to do in order to improve this file-based collection and its data, both to ensure its preservation and to provide access as part of the wider archive landscape in the BBC: the systems will likely always continue to evolve. Evolution will soon be taking place as part of the LTO3 migration even before ingest of the current batch of videotapes is complete – the use of commonly adopted and well-supported open standards may ease future migrations of the collection.

This paper has only considered part of the videotape collection. The remaining videotapes and other carriers (e.g. film) around the organisation will eventually need to be processed, perhaps using different systems and file / storage schemes. Of course, the collections of radio programmes, photos, documents (contracts, scripts, etc) are all other challenges for the BBC, all at different 'levels' of digitisation and digital preservation.

6. REFERENCES

[1] Glanville, M. and Heritage, T. 2013. *A Guide to Understanding BBC Archive MXF Files.* White Paper 241. BBC R&D. www.bbc.co.uk/rd/publications/whitepaper241

[2] Wright, R. 2012. *Preserving Moving Pictures and Sound.* Digital Preservation Coalition. http://www.dpconline.org/component/docman/doc_download/753-dpctw12-01pdf

[3] *Ingex.* BBC R&D. http://ingex.sourceforge.net

[4] Easterbrook, J. *The BBC Transform PAL Decoder.* http://www.jim-easterbrook.me.uk/pal/

[5] *Reference Model For An Open Archival Information System*. CCSDS 650.0-M-2. 2012. http://public.ccsds.org/publications/archive/650x0m2.pdf

[6] Wright, R. and Addis, M. 2010. *Audiovisual Preservation Strategies, Data Models and Value-Chains.* PrestoPRIME / PrestoCentre. http://www.prestocentre.org/library/resources/audiovisual-preservation-strategies-data-models-and-value-chains

[7] *Audit and Certification of Trustworthy Digital Repositories*. CCSDS 652.0-M-1. 2011. http://public.ccsds.org/publications/archive/652x0m1.pdf

[8] *Digital Master Archive Format.* Forum discussion. PrestoCentre. http://www.prestocentre.org/forum/digital-master-archive-format

[9] Owens, T. 2012. *NDSA Levels of Digital Preservation: Release Candidate One.* Blog post. Library of Congress. http://blogs.loc.gov/digitalpreservation/2012/11/ndsa-levels-of-digital-preservation-release-candidate-one/

[10] *Digital Archiving Maturity Model.* Tessella. 2012. http://www.digital-preservation.com/wp-content/uploads/Maturity-Model-Web.pdf

[11] *Digital Preservation Management.* Tutorial. http://www.dpworkshop.org/dpm-eng/conclusion.html

[12] Addis, M., Jacyno M. et al. 2012. *Tools for Quantitative Comparison of Preservation Strategies.* PrestoPRIME. http://prestoprimews.ina.fr/public/deliverables/PP_WP2_D2.1.4_PreservationStrategyTools_R1_v1.00.pdf/view

[13] *Presto4U Project: Work Packages.* http://www.prestocentre.org/4u/work-packages

[14] *PREMIS Data Dictionary for Preservation Metadata.* Version 2.2. 2012. http://www.loc.gov/standards/premis/v2/premis-2-2.pdf

[15] *MPEG-A: Multimedia Preservation Application Format* http://mpeg.chiariglione.org/standards/mpeg-a

[16] *MXF Archiving & Preservation.* AMWA AS-07. http://www.amwa.tv/projects/AS-07.shtml

[17] Heritage, T. 2011. *Archive Packages: Containers for Complexity.* FIAT / IFTA World Conference (Turin, Italy, September 2011). http://fiatifta.org

[18] Phillips, S. 2010. *Service Level Agreements for Storage and Preservation.* PrestoPRIME / PrestoCentre. http://www.prestocentre.org/library/resources/service-level-agreements-storage-and-preservation

Large-Scale Curation and Presentation of CD-ROM Art

Dragan Espenschied
Karlsruhe University of Arts and Design
Lorenzstr. 15
76135 Karlsruhe, Germany
dragan.espenschied@hfg.edu

Klaus Rechert
University of Freiburg
Hermann-Herder Str. 10
79104 Freiburg i. B., Germany
klaus.rechert@rz.uni-freiburg.de

Dirk von Suchodoletz
University of Freiburg
Hermann-Herder Str. 10
79104 Freiburg i. B., Germany
dirk.von.suchodoletz@rz.uni-freiburg.de

Isgandar Valizada
University of Freiburg
Hermann-Herder Str. 10
79104 Freiburg i. B., Germany
isgandar.valizada@rz.uni-freiburg.de

Nick Russler
University of Freiburg
Hermann-Herder Str. 10
79104 Freiburg i. B., Germany
nick.russler@googlemail.com

ABSTRACT

For memory institutions both preservation and presentation of digital art is especially challenging. The digital toolset available to artists is almost infinite as well as their creativity using technology in unconventional ways. In contrast to other sources of digital artifacts with sector-wide quasi-standards on digital formats, each artwork presents a challenge of its own. Hence, each object need individual examination and preparation in order to preserve it in a useful way.

In this paper we present workflows and tools for emulation-based preservation and presentation of digital art by the example of a collection of CD-ROM art. Furthermore, we evaluate the performance results of an emulation-based approach.

Categories and Subject Descriptors

H.3 [**Information Storage and Retrieval**]: Digital Libraries

General Terms

Digital Preservation, Digital Art, Long-Term Access, Emulation, Performance, Authenticity

1. INTRODUCTION

Preservation of digital art poses new challenges to memory institutions, both with respect to curation and presentation. The objects to be preserved consist of dynamic, interactive artifacts designed for computer systems of their time. As these artworks rely on media and platforms with life-cycles of less than a decade, new ways for preserving access are required. In contrast to simple digital artifacts like text-based documents, digital art can be very challenging w.r.t. technology. The digital toolset available to artists is almost infinite as well as their creativity using technology in unconventional ways. Furthermore, there are no discipline related format standards, preventing a generalized approach. Moreover, digital art artifacts cannot be migrated to formats that are easier to maintain (i.e. video) without losing their interactive performance. In many cases, there is no clear distinction between an "interaction" and "content" that is interacted with.

Thus, memory institutions require versatile strategies to preserve, curate and display digital art efficiently. Emulation technology is able to provide a base technology for this task, for instance, to keep digital artifacts alive. However, having suitable emulators and related technology is generally not sufficient. A framework, i.e. integration of archives and repositories, workflows and best-practices are required to cope with today's and upcoming challenges. In this paper we show the adaption of the bwFLA framework to integrate tools and workflows to curate and present a large and challenging collection of contemporary digital art.

2. DIGITAL CULTURE – MASS CULTURE

A lot of digital art should be easily accessible, without too much emphasis on traditional aura and exclusivity. Like digital culture did in many areas, wide availability of tools and constant change in technology and theory made it an attractive entrance into the art world for newcomers and young artists. In the field of digital art, the general attitude of most participants is that anyone is always welcome to join in and spur the discourse. When it comes to longevity, however, it is quite difficult to find a suitable place for the resulting amount of artworks to survive in the swiftly changing technological landscape.

This is an imbalance that needs to be tackled if digital art and digital culture as a whole should be able to create a notion of history: artists are having difficulties building a recognized body of work, institutions are having difficulties building a reputation. As a result, many mid-career artists will turn to more durable and therefore, sellable objects and formats.[1]

[1]Marius Watz, one of the few media artists daring to speak about economic conditions and has moved away from producing for digital displays into producing sculptures, is linking the precarious situation of many of his colleagues to a lack of history in his "provocations":

> The success of media art is NOT a matter of time. Media art history is constantly being forgotten. [10, p30]

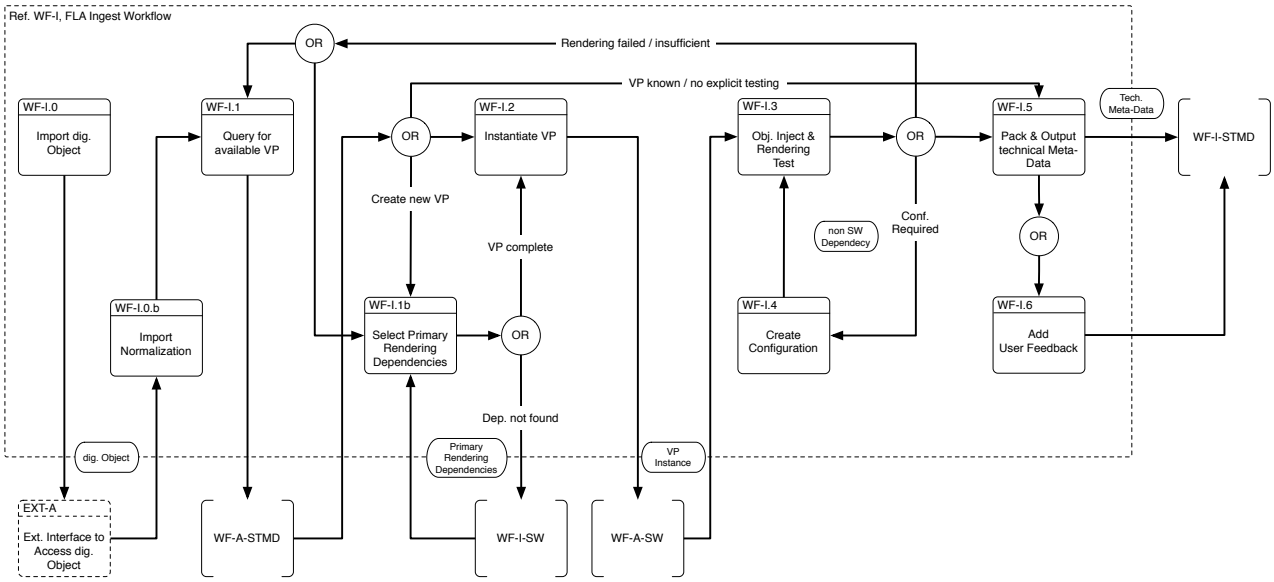


Figure 1: bwFLA: Ingest conceptual model.

What is left over of important periods and movements of digital culture is often no more than printed screenshots in art books and exhibition catalogues. In some cases, they are looking good on paper, but access to the actual artifacts is unavailable. There is no way to actually compare or re-evaluate these artifacts a second time, discuss their relevance retrospectively, analyze interactions in between them or just keep the discussion around them going – and this quite drastic cultural cut-off happens with objects that are just a few years old.

A useful and quite promising approach to curate highly volatile digital art is recognizing it as mass culture, which in turn requires a mass-curation approach. In this case, mass doesn't mean uniformity and is not referring a broadcasting model of media, but mass authorship and mass prosuming. This understanding makes it possible to define more abstract "significant properties" on a large collection of artifacts.

Digital culture is practices rather than artifacts.

No digital artifact, regardless of its manifestation, e.g. as file, an executable, a memory-dump, or similar, carries a history within itself. One of the main features of digital data is that any part of it can be easily changed, re-contextualized or removed without a trace. Of course, the same applies for digital art.

Knowledge is mostly embedded in practices. History is comprehended as the understanding of how and in which contexts a certain artifact was created and manipulated and how it affected its users and surrounding objects. For instance, the process of how a web page is built and the almost infinite amount of technological environments in which this artifact can be used (browsers, image manipulation software, text-editors, word processors, video editors, printers) for different purposes and with different motivations (rearrangement, comparison, re-design, plagiarism, collecting, entertainment) is crucial for understanding and classifying the artifact.

For a meaningful preserving of such artifacts, this means that a memory institution needs to provide methods of interaction and manipulation for its collection. Otherwise it will be impossible to make sense of them outside of speculation.

Authenticity does not scale.

In theory it might be possible to reconstruct environments for almost any single digital artifact that re-enacts its performance exactly "as the artist intended" given suitable (financial) resources. In general, however, such an approach is either inefficient, i.e. too laborious for many artifacts, or it makes no sense because the artist's specifications cannot be met technically or logistically. Finally, the whole idea of individual technical restoration may not match the artifact's main performance feature because it unfolds its impact in mass usage and distribution, and therefore has no "form" outside of practice.

Since there is no single way to render, view or use a digital artifact, it is futile attempting to define one. There might be even no time to read every artist's statement on how an artifact should be handled. — The information contained in "installation instructions" might as well be considered meaningless: Detailed instructions are typically defined for special situations like exhibitions, but have no effect on the artifact's behavior outside of them. While artists can define what type of monitor or projection shall be used to display a work in a museum or gallery, they have no say in how and when their work is accessed by for example web users, they cannot even dictate and ephemeral nature of their creation.

Instead, making the largest possible amount of artworks accessible in combination with providing broadly generalized forms for their interaction and manipulation, seems like the most worthwhile approach. The outcome is a reduced amount of rich simulated environments that enable the interaction of and with more artifacts.

This is not about disrespecting individual artists, but to generally **enable** discussions about certain forms of art and artists. Hence, in order to create the possibility for artifacts to reach cultural and historical significance in the first place, fidelity and ease of access need to be balanced. What "ease of access" means depends on technology and usage patterns available at the time of access. In general, the least expert knowledge is needed to interact with an artifact, the better. The highest grade of accessibility is the possibility for general users to be confronted with an artifact and interact with it in their typical context. Example: screenshots of interactive works are easy to distribute, post on social media sites, archive, modify. The accessibility of an re-enactment can quite easily be asserted, fidelity, though, is an open-ended scale.

Poetic Qualities of Emulation and Digital Art.

It is meaningful to not only rely on emulation to deliver a historic performance, but to develop expressive devices on top of emulation that can serve as building blocks for environments. Different output devices and some of their glitches can be staged, their effects combined. For instance, simulating the image structures of low-resolution CRT screens on today's high-resolution LCD screens is a common technique used in the video game emulation community. While there are different CRT software emulators available and some enthusiasts are working to thoroughly replicate properties of precise monitor models in software, users are usually free to choose which display mode looks best for them.

Preservation of digital art must build upon this model for re-enactments in order to address the infinite context problem: It might be impossible to deliver the exact performance of an historic monitor, but a simple CRT fake that can be switched on at will might enhance the performance of many artifacts at once. Similarly, it is unfeasible to connect a snapshot of the whole Internet at a certain point in time with a historic artifact from that time. Yet, a "good enough" fake that can mimic popular services up to a certain point of interaction based on a simple archive can be developed and provided to enhance the performance of a whole class of artifacts. It is not even necessary to define these classes in advance, as an improvement of the framework might affect an unknown number of artifacts' performances.

Artifacts that either rely on certain subtleties of their environment that are hard or impossible to cover via emulation or staging, or are requiring a context too large to re-create or stage, are certainly losing some of their quality. Both of these types of work are actually easy to create for artists. Once the re-enactment of historic digital art is established as a task spanning more than one piece, but rather whole genres, periods and movements, it will be possible to approximate even these cases to an agreeable quality. Already before, each working artifact carries the possibility to enhance the performance of every other artifact.

3. RELATED WORK

Geoffrey Brown and colleagues also tackled the problem of preserving CD-ROMs as well as providing access by using an emulation-based strategy [11, 1]. To enable several institutions making use of and potentially contribute to the collection, the CD-ROMs are served through a distributed filesystem. Further, they require some client preparation regarding emulator setup. Compared to local provisioning of a complex service stack as also proposed in KEEP [6, 5], a networked approach reduces technical and organizational hurdles at the client's side significantly. In contrast, we present a versatile server-based infrastructure providing functional access to a wide range of emulators and operating systems without any requirements regarding the user's client environment beside a standard modern web browser.

With respect to authentic preservation and presentation of complex digital objects, in particular digital art, the discussion can be divided into a technical and an art-related part. Guttenbrunner et al. provide a generic framework qualifying emulator performance [3]. Furthermore, there is a lively discussion on authentic simulation of individual technical components such as CRT screen simulation [9, 2]. Following the discussion above, the goal of the bwFLA framework, but also the focus of this work, is providing convenient access to current emulator technology and corresponding digital objects. By using an emulation-as-a-service architecture model, new emulators and technology can be integrated with reasonable effort, being then available for any already present digital artifact. Regarding the preservation of digital art Perla Innocenti also pointed out the difficult notion of authenticity in this context [4]. She introduced the concept of dynamic authenticity, also proposing a variable approach and object-centric approach, allowing a certain degree of tolerance "to match digital art intrinsic variability." Similar, in this work we focus on a pragmatic approach by today's available emulator technology.

4. ENVIRONMENT AND TOOLS

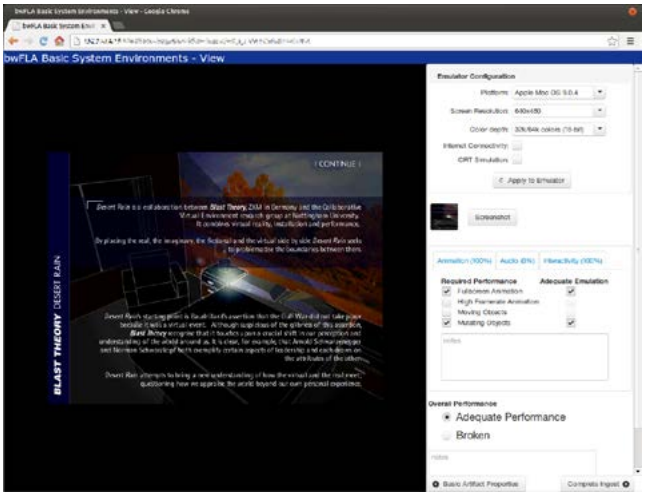
The Baden-Württemberg Functional Long-Term Archiving and Access (bwFLA) is a two-year state funded project transporting the results of past and ongoing digital preservation research into practitioners communities. Primarily, bwFLA creates tools and workflows to ensure long-term access to digital cultural and scientific assets held by the state's university libraries and archives. The project consortium brings together partners across the state, involving people from university libraries computing centers, libraries and archives providing a broad range of background and insights into the digital preservation landscape.

Workflows and tools developed by the bwFLA project are designed to be used in a distributed, labor- and cost-sharing setting. While the project delivers technical solutions and a distributed service-model, preservation of individual digital objects and accompanying measures are left to individual memory institutions. The goal of the bwFLA framework is to enable these institutions to use tools and perform workflows on certain types of digital objects, both for ingest and access workflows.

The bwFLA Emulation-as-a-Service (EaaS) framework and its abstract workflows have been adapted to a specific domain, in particular a CD-ROM digital art collection. Technical background of the emulation framework can be found in earlier work [8]. This paper focuses on providing tools and an actual workflow implementation

- to quickly look through a large collection of yet unknown digital objects,
- to create technical meta-data describing a working rendering environment,

(a) Ingest Step 1: Selecting a CD from the archive.



(b) Ingest Step 2: Rendering and evaluation of a digital artifact.

Figure 2: bwFLA: Ingest workflow implementation.

- to create meta-data describing content features,
- to organize the collection by creating screenshots and videos,
- to provide a simple access platform for a general audience.

For this, the framework provides three basic workflows: ingest, preparation of rendering environments and access. In this paper we will discuss ingest and access workflows in detail, while a detailed description of the environment preparation workflow is given in earlier work [7].

4.1 Ingest

The bwFLA ingest workflow is designed as a flexible and optional extension of traditional ingest workflows. Therefore, we assume that basic archival meta-data is already recorded and available, i.e. due to a previous basic archival ingest workflow. Similar, we assume that the digital object is available through some archival identifier and can be retrieved through a dedicated interface. Starting from the conceptual model depicted in Fig. 1 a specific workflow instance to describe CD-ROM artwork has been adapted and implemented.

As a first step of the emulation ingest workflow, the manifestation of the digital object is normalized (WF-I.0 esp. WF-I.0.b). In our case, we have received a CD-ROM collection containing either a folder consisting of an ISO file together with a thumbnail image and in some cases a description or we have received directories containing all CD items as individual files. For the latter case, ISO files were created as part of the workflow normalization step.

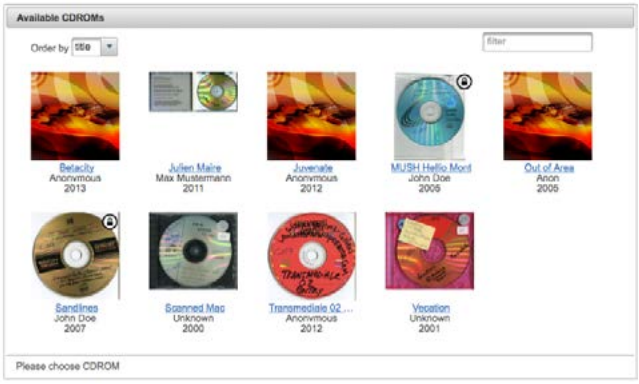
The user is then able to select an individual object by requesting a specific archival ID. Additionally available information from the archival meta-data is displayed as reference. If this data is incomplete, the user is able to complete the data-set. To keep the workflow as simple as possible, only two types of rendering environments are selectable: Apple Macintosh and Microsoft Windows. The concrete operating system version is chosen automatically based on the object's production year. In case of special requirements, the user is able to run a system image preparation workflow to create a specialized rendering environment (denoted as WF-I-SW in the conceptual model). Fig. 2(a) shows the correspondent user interface.

At this point, an emulation component has been allocated and set up and the object has been prepared to be injected into the rendering environment. With WF-I.3, the second phase of the ingest process begins with starting the chosen rendering environment and injection of the digital object. At this point, the user is required to evaluate the quality of the object presentation. Next to the emulator output (cf. Fig. 2(b)), the user is able to describe technical configuration details (such as optimal screen resolution, color depth etc.) and the object's desired/expected as well as the actual performance in the chosen environment. Gathering this feedback is used to compile a fidelity rating that is displayed during access and may help users to interpret imperfect emulation results or to choose only artifacts that are emulated in a certain quality. This meta-data set is rather domain-specific and discussed in detail in Section 5. At this step, also auxiliary material such as screenshots or video captures, could be produced, e.g. to enrich catalog records.

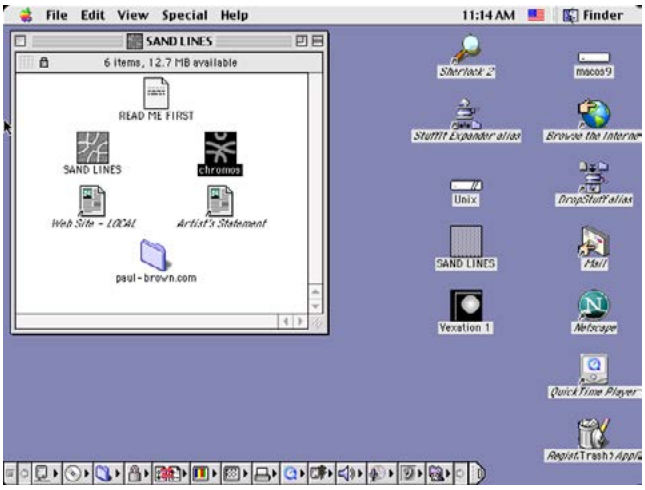
As a result of the bwFLA ingest workflow, technical meta-data is created describing both the technical setup, i.e. technical details rendering the environment's view path, user configuration as well as domain specific performance measures. The data is encapsulated either as JSON object or XML file and is delivered to the object owner. Archiving of meta-data as well as the objects are the owner's responsibilities since bwFLA only provides the technical framework. Only a limited set of rendering environments are kept, to improve usability and access to objects. With the provided technical meta-data, however, an automatic or semi-automatic instantiation of a view-path is available as part of the WF-A-SW workflow [7].

4.2 Access

Having appropriate meta-data, the bwFLA access workflow provides convenient access to archived digital objects.



(a) Available CD-ROMs.



(b) Digital object rendered in emulated environment.

Figure 3: bwFLA: Functional access to digital art collection.

Similar to the assumptions made at ingest time, we assume that objects and meta-data are accessible through a dedicated networked interface. The user is then able to choose an object from the catalog, browse its meta-information and finally start the rendering process. Fig. 3(a) shows the current bwFLA catalog implemented for the Transmediale festival [2] collection. Based on the provided meta-data the requested environment is instantiated and started with the digital object attached. Fig. 3(b) shows the rendered result.

5. QUANTITATIVE ANALYSIS

As a test-case, the Transmediale archive kindly provided us with their collection of CD-ROM art in the form of ISO or bin image files. Most of the objects were created in between 1995 and 2005, the largest part during the height of the genre around 1999 and 2001.

The Transmediale's goal is to make this collection publicly available online. Since the collection contains 272 pieces, it is unfeasible to analyze each one in-depth. Instead, a representative selection of six CD-ROMs was picked out to be evaluated on original, consumer grade Apple hardware that was also often used in exhibition settings. This performance was compared to the performance on a stock emulator. Based on this analysis, a very limited list of performance properties was created that only consider the emulator's performance over the network and configuration.

Technical requirements like processor speed, amount of RAM and exact OS version are not very important in the context of CD-ROM art, and usually available as part of the provided view-path. If the artist had specified certain setups, in most of the cases this information turns out to just be the artist's own setup used to create the work. Since most CD-ROM art was created using Macromedia Director (an integrated authoring software catering to the lowest technical dominator), great performance issues are not to be expected. The operating system version to use can be extrapolated from each CD-ROMs publishing date. For instance, an exact Quicktime version to replay a video is even less critical, as Quicktime was sufficiently backwards compatible: newer versions could replay all older versions' videos. As the Quicktime library was used to embed videos into Macromedia Director, how the actual player controls look in different versions is irrelevant because they are never visible anyway.

[2]Transmediale Festival, `http://www.transmediale.de/`

Technical features & User configuration.

The first subset of meta-data collected describes objectively perceptual technical features and user-configuration. A view-path is able to describe the general technical setup starting with a detailed description of the emulated hardware to installed application, fonts and libraries.

However, many visible features of a system environment depend on individual user-configuration:

- **Platform:** This property describes a suitable rendering platform for a digital object. Usually the platform is named after its backing OS and or hardware. The description of the rendering platform describes the combination of a concrete software stack (view-path) and its specific emulator configuration.
- **Network access** might be required for an artifact or may enhance its performance. This property describes if a network setup is provided by the chosen platform, i.e. the emulator provides suitable networking features and the chosen platform is configured properly.
- Vision and Sound: **Pixel resolution** and **color depth** have an impact on the computational performance of the emulator as well as the infrastructure that is required to deliver its results to the user. Especially interactive works can benefit greatly from a possible reduction of system reaction time. While this property is a technical feature it is usually only observable if the specific platform is instantiated.
- Information about **animation quality** enables to create an educated compromise setup about interactivity, pixel resolution and color depth. In some cases, a smooth movement of on-screen objects is more important for an artifact's performance than accurate color

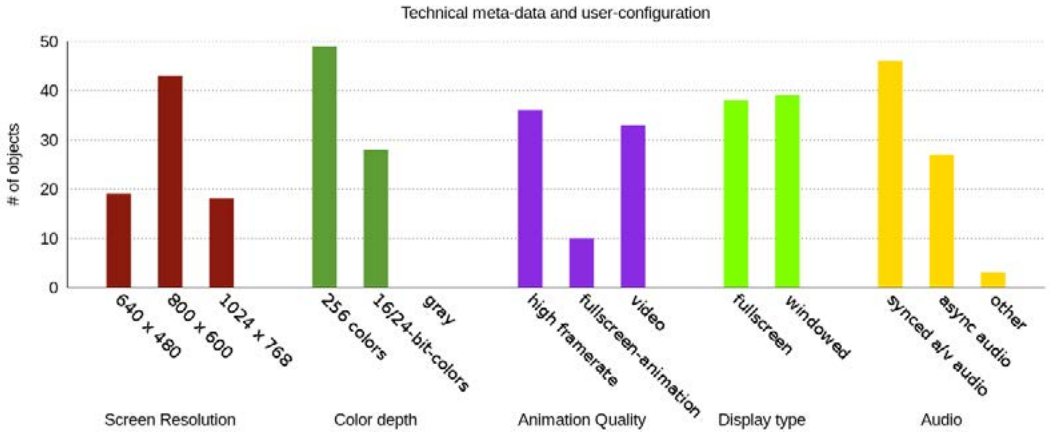


Figure 4: Preliminary results: technical features and user-configuration.

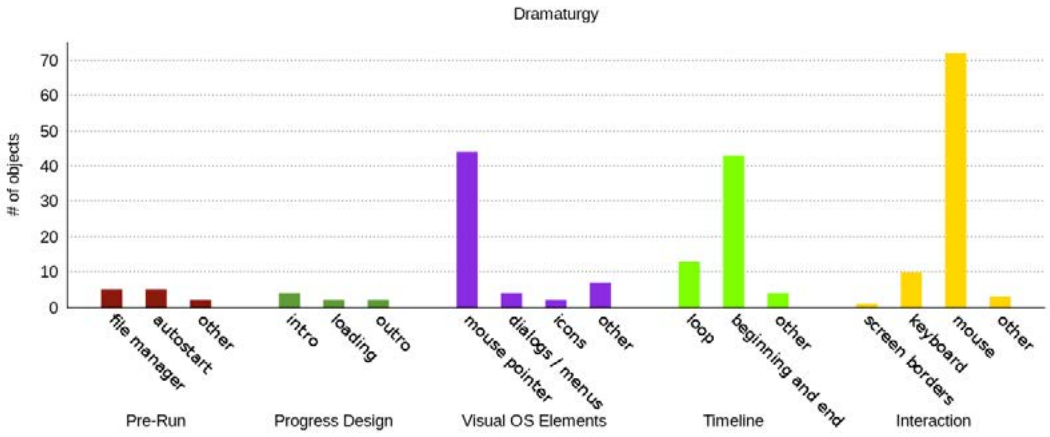


Figure 5: Preliminary results: dramaturgy of presentation.

representation or detail, others might rely more on vibrant still images. Full screen animation, where each animation frame replaces the contents of the complete frame buffer, are especially challenging.

This is an important criterion because the general user base is very aware of "snappy" versus "sluggish" animation. Consumer devices have been advertised on the basis that their reaction times are very low, there are established terms and language to describe different rates of performance.

High and low color depths and resolutions are also easily distinguished by users because low-color, low-resolution graphics are accepted as form of expression of its own and are often associated with "old computers". Since the collection of CD-ROM Art is explicitly "old", a reduction of color depth will be more accepted by the audience than a loss of animation quality.

- **Synchronicity**: Are audio and video in any way related to each other? Some CD-ROMs might have a sound track that is not in sync with what is visible on screen, others trigger sound effects to emphasize certain visual events or interactions. Disconnected audio loops can be delivered to the user on a side channel, for example a separately downloadable audio file. The emulator would not need to take care of sound in this case at all. In the synchronous case, the sound effects' bit-depth and sample rate might be needed to be sacrificed. Since audio is used in so many contexts (headphones, active and passive speakers of different quality, mobile phone speakers, etc), users usually cannot notice or describe the bit depth or general fidelity of sound without two comparable recordings being replayed.

  Well-synced audio and video makes a great user experience and should always be considered more important than sample rate, bit depth and color depth.

Dramaturgy.

A sub-set of domain specific meta-data, describing expectations on the artwork's general performance and dramaturgy.

- **Structuring of time**: How a CD-ROM is making use of dramatic devices over time, can greatly affect the emulator's configuration. If the work makes no dramaturgic reference to the underlying operating system

and does nothing during loading, or even can run as a loop, the emulator can be set up to launch straight into the software without presenting any parts of the operating system to the user. This might greatly reduce the effort of constructing an emulation environment.

- **Interactivity**: If the work requires any kind of user interaction to perform, the emulator has to provide adequate facilities. If the work is not interactive, there is no need to provide those.

6. EVALUATION RESULTS

For finding the evaluation criteria, six artifacts were examined in-depth. At the moment of this writing, the evaluation on original hardware is still ongoing, 86 of 272 artifacts have been checked in the course of only a few days.

It was found out that only 5 artifacts make use of a sophisticated arrangement of icons and windows in the operating system environment. Only 4 artifacts made use of a designed intro sequence, 2 artifacts featured a designed loading sequence. Apart from the operating system's mouse pointer images, which were appearing in 44 pieces, no visual elements of the operating system were used within interactive pieces. 72 pieces can be considered interactive, 43 have a beginning and an end, only 13 can run in a loop. 38 run in full screen and thereby hide the underlying operating system completely. Figure 4 and 5 visualize evaluated features of our preliminary results.

The genre of CD-ROM art didn't seem to be too reflective of or alluding to its software environment. Ultimately, it is not even reflective of the CD-ROM and its media specific properties, like slow loading and read-only data access. Instead, authors seemed to strive for closed, narrative, interactive experiences. For the access part, this means that the boot-up process of the emulated operating system and even the loading of the CD-ROM's data can be skipped over and instead a saved state of the emulator with the work already fully prepared to be interacted with should be presented to the user. Witnessing the preparation process should be the user's choice at the moment of access. In order to define a saved state, the ingest process has to provide means of "freezing" a state of the emulator, in bwFLA this is implemented with a "save current state" button in the evaluation GUI.

33 works apply digital video replay, 46 require synchronous audio and video. In many cases, the sound effects are simple clicks to re-assure the user of an interaction acknowledged by the system – however, if these sound effects are delayed, they can cause a lot of confusion. In these cases, it would be better to turn off sound in case the emulator or network infrastructure cannot produce synchronicity. Again, during access, the user should be presented with the option to turn off sound.

7. LESSONS LEARNED

While the used emulators themselves (QEMU[3], SheepShaver and BasiliskII[4] do not have problems running any of the artifacts with adequate performance locally, the networking layer and standardized browser clients introduced by the EaaS approach are responsible for the widest variety in performance. While the networked setup of bwFLA's EaaS approach reduces the technical hurdles using emulation significantly, latency and sound transport issues may reduce the overall performance results.

As a result, it makes sense to collect single evaluation results per artifact from tests on original hardware and multiple evaluation results per artifact from tests on the emulator. The single tests on original hardware are mainly serving the purpose to define what properties are to be expected and should be checked later in the emulator. The multiple evaluation results per artifact contain performance connected together with certain values describing the state of the emulator and the client at runtime (emulator setup, load of the emulation host, network throughput, network delay, load of the client, etc). This allows to compare different setups and create an estimation of fidelity for future runs, to point users' attention to properties that might not perform quite as they should. Users then can adjust their expectations and perception of the artifact's performance, leading to more conscious interpretations. Multiple evaluations also open up the possibility of a community-based evaluation, where users with knowledge of a certain detail in the original performance can give feedback on how well the emulator performance matches, or can create different, alternative EaaS set-ups that improve a certain aspect or all of the performance of an artifact.

During evaluation of the emulator's performance, it has proven more practical to ask the user only to check for failures of the system, as they are more apparent to identify: jerky animation due to network latency is obvious to spot, manually running through lists of always the same properties that apparently work fine and confirming each seems like a waste of time.

8. DISCUSSION & CONCLUSION

Digital art is a new challenge for memory institutions and galleries, which is different from traditional art forms, even from the newer developments like video art. Simple screenshots or video recordings cannot capture all object properties. The presented method supports preservation of digital art, especially for object appraisal and curation. The method can be easily extended to other uses, like the presentation of non-standard, interactive artifacts in libraries. The same applies for the cataloging and curation of a wide range of computer games and can be the tool of choice to sift through software archives.

Some issues, however, remain open at this point. Most urgently, licensing, especially with regards to distributed architectures such as bwFLA's EaaS model, is a huge hurdle. While such a model contributes to usability and enables access to digital art for a wide audience, the legal situation especially, regarding operating system and software, needs more attention, since this leads to a paradox situation. While ancient operating systems and software packages have lost most probably their commercial value as well as their functional utility, digital art most probably increases its cultural level with time progressing.

[3]Open Source Processor, `http://wiki.qemu.org/Main_Page` (version of 6/28/2013)

[4]BasiliskII and Sheepshaver, Open Source M68K and PowerMac Emulator, `http://sheepshaver.cebix.net/` (version of 6/28/2013)

Acknowledgments

The work presented in this publication is a part of the *bwFLA – Functional Long-Term Access*[5] project sponsored by the federal state of Baden-Württemberg, Germany.

9. REFERENCES

[1] G. Brown. Developing virtual cd-rom collections: The voyager company publications. *International Journal of Digital Curation*, 7(2):3–22, 2012.

[2] M. Guttenbrunner and A. Rauber. Re-awakening the philips videopac: From an old tape to a vintage feeling on a modern screen. In *Proceedings of the 8th International Conference on Preservation of Digital Objects (iPres 2011)*, pages 250–251, 11 2011. Posterpresentation: iPres 2011 - 8th International Conference on Preservation of Digital Objects.

[3] M. Guttenbrunner and A. Rauber. A measurement framework for evaluating emulators for digital preservation. *ACM Trans. Inf. Syst.*, 30(2):14:1–14:28, May 2012.

[4] P. Innocenti. Rethinking authenticity in digital art preservation. In *9th International Conference on Preservation of Digital Objects (iPRES2012)*, pages 63–67. University of Toronto, 2012.

[5] B. Lohman, B. Kiers, D. Michel, and J. van der Hoeven. Emulation as a business solution: The emulation framework. In *8th International Conference on Preservation of Digital Objects (iPRES2011)*, pages 425–428. National Library Board Singapore and Nanyang Technology University, 2011.

[6] D. Pinchbeck, D. Anderson, J. Delve, G. Alemu, A. Ciuffreda, and A. Lange. Emulation as a strategy for the preservation of games: the keep project. In *DiGRA 2009 – Breaking New Ground: Innovation in Games, Play, Practice and Theory*, 2009.

[7] K. Rechert, I. Valizada, and D. von Suchodoletz. Future-proof preservation of complex software environments. In *Proceedings of the 9th International Conference on Preservation of Digital Objects (iPRES2012)*, pages 179–183. University of Toronto Faculty of Information, 2012.

[8] K. Rechert, I. Valizada, D. von Suchodoletz, and J. Latocha. bwFLA – a functional approach to digital preservation. *PIK – Praxis der Informationsverarbeitung und Kommunikation*, 35(4):259–267, 2012.

[9] J. Scott. What a wonder is a terrible monitor. Online `http://ascii.textfiles.com/archives/3786`, 2012.

[10] M. Watz. The futility of media art in a contemporary art world. Online `http://www.scribd.com/doc/110282792/20121015-LISA-The-Futility-of-Media-Art-in-a-\Contemporary-Art-World`, Oct 2012.

[11] K. Woods and G. Brown. Assisted emulation for legacy executables. *International Journal of Digital Curation*, 5(1), 2010.

[5] bwFLA – Functional Long-Term Access, `http://bw-fla.uni-freiburg.de`.

Interoperability Objectives and Approaches: Results from the APARSEN NoE

Barbara Bazzanella
Department of Information Engineering and Computer Science (DISI), University of Trento, Italy
barbara.bazzanella@unitn.it

Yannis Tzitzikas
Institute of Computer Science, FORTH-ICS, and Computer Science Department, University of Crete, Greece
tzitzik@ics.forth.gr

ABSTRACT

In this paper we report the main results of a study on interoperability objectives and approaches in digital preservation (DP), conducted within the APARSEN Network of Excellence (NoE)[1]. The aim of the investigation was to collect interoperability challenges and goals from various initiatives and project partners and to produce a matrix of solutions and guidelines that can guide the stakeholders in DP to the multi-dimensional and complex landscape of digital preservation interoperability. The paper describes the main findings of the research, including 1) an overview of the current projects and initiatives on interoperability in different areas of digital preservation, 2) an analysis of the main interoperability scenarios and challenges encountered by partners and other stakeholders in their daily life activity that served to drive the definition of the main common interoperability objectives for digital preservation, 3) a broad matrix of models, standards and services for interoperability that cover the main areas of digital preservation, which can be used as a working instrument to navigate the complex ecosystem of the current interoperability solutions, and 4) a list of recommendations and guidelines to create the ground for a coordinated and interoperable digital preservation ecosystem.

1. INTRODUCTION

Interoperability refers to the ability of two or more independent systems to exchange information and use the exchanged information in meaningful ways and without special effort to achieve common goals [4, 1]. Interoperability has become a critical imperative for digital preservation in recent years and several initiatives have started to focus on the definition of requirements, technological solutions and best practices in order to define digital preservation interoperability frameworks, services and standards for effectively and reliably access the preserved digital content between interoperating systems. This shows the general agreement within the DP community that an effective DP strategy or solution strictly relies on a broad international consensus on interoperability, as well as on appropriately designed technological infrastructures to enable it. Identifying the interoperability issues involved and the interoperability objectives to achieve is a first step to promote such a consensus. However this is not a trivial task due to a number of aspects to consider. On the one hand, digital preservation has started to be approached as a problem of "interoperability with the future" [10] or "temporal interoperability" [5], that is ensuring that current systems interoperate with future systems to guarantee that digital resources remain accessible and reusable over a long period of time maintaining their meaning and value. According to this definition, the techniques used for contemporaneous interoperability are applicable for temporal interoperability (i.e. digital preservation), indicating many potential commonalities and points of synergy between interoperability in real time and digital preservation even though temporal interoperability requires a specific focus on sustainability and applicability of the same strategies in the long term. On the other hand, the resources that need to be preserved are highly heterogeneous and increasingly distributed across different systems and organizations which should interoperate in real time, share responsibilities and rely on each other to provide integrated and cross-boundary DP services. The temporal dimension of interoperability is just one aspect of the complexity of the interoperability landscape in DP. First of all, interoperability is a very broad and complex concept, which is conceived on different levels of abstraction (as discussed in the next section) ranging from syntactic to semantic interoperability [7] passing through technical, political, organizational and legal perspectives [8] and dealing with many interoperability objects (e.g. metadata, persistent identifiers, policies). Secondly, several interoperability issues cut across different areas of digital preservation (e.g. Persistent Identifiers, Authenticity and Provenance, Preservation services) showing a very fragmented landscape where there is relatively little harmonization of models, standards and services used in the creation, management and preservation of digital cultural contents. Finally, different stakeholder communities deal with a broad range of interoperability challenges and barriers, which affect in many ways different local functionalities and approaches.

Diagnosing this complex ecosystem is a first fundamental step in order to reach a common awareness about the main interoperability challenges in DP and to define a core set of interoperability objectives for the future. The NoE of the APARSEN project should play a key role to coordinate the definition of this agenda due to its commitment in the creation of a common view and understanding about the preservation and interoperability requirements in different preservation domains, communities and research areas. This paper aims at providing a contribution in this context, summarizing the main results of an investigation on interoperability objectives and approaches conducted within the APARSEN project. First of all, it gives a broad overview

[1] http://www.alliancepermanentaccess.org/index.php/aparsen/

of ongoing and past projects and initiatives covering interoperability issues related to digital preservation. Secondly, the paper discusses interoperability scenarios and challenges encountered by partners and other stakeholders. Third, a broad matrix of interoperability models, standards and services is described as a working tool to navigate the complex interoperability ecosystem. The paper closes with an initial set of recommendations which should promote the realization of an interoperable long-term preservation ecosystem. More details and results can be found in the public deliverable, D25.1[2].

2. WHY INTEROPERABILITY IS IMPORTANT FOR DIGITAL PRESERVATION

A study conducted by the EC in 2011 mentions "interoperability" as one of the ten most important research topics for digital preservation research[3]. In this section we discuss why addressing digital preservation issues with a focus on interoperability may offer significant advantages over current practices for ensuring access, exchange and reuse of digital content in the long term.

First of all, digital preservation certainly requires preserving the bits of the digital objects, but this is probably the less difficult task. The preservation of their accessibility, intelligibility, provenance, authenticity, quality (and many others, e.g. citability, searchability, etc) is a more complex task. All these requirements can be considered as interoperability aspects, in the sense that they can be considered as abilities to apply (now and in the future) successfully in different objects the same operations for accessing them, understanding them, rendering them, getting their provenance information, etc. This is why digital preservation has been termed "interoperability with the future". Moreover, interoperability usually refers to the ability to "exchange and use information between independent systems in meaningful ways and without special effort". As a consequence, achieving interoperability (according to this definition), implies ability to exchange and use information without special effort, thus preservation of accessibility, intelligibility, etc, without special effort.

Secondly, expressing crucial digital preservation challenges as interoperability challenges has a beneficial impact not only for the design and implementation of scalable technical solutions, but also for the definition of a common research agenda agreed by stakeholders, which are concerned with long-term preservation and stakeholders that are focused on building interoperable digital environments. By recognizing that common needs and issues are in play, it should be easier to adopt integrated solutions and expand the applicability of standards and models developed within a certain context to data created and used by other communities and across technical, organizational, political and social boundaries.

Third, DP can be conceived as an interoperability exercise along the entire spectrum of steps that form the lifecycle of a digital object, from its creation to its re-use through the process of preservation. A fundamental aspect of this exercise is the adherence to digital preservation standards, as pointed out by (National Information Standards Organization, 2004) "An institution must ensure that its standards are in line with those used across the digital library community to enable interoperability where possible". To this purpose, digital preservation standards should not be conceived from a repository-centric point of view but should be defined as a set of functional requirements which can be implemented by multiple systems with different hardware and software platforms, data structures, and interfaces to manage and exchange data in the medium and long term with minimal loss of content and functionality.

Finally, in the global context of digital information, DP has more and more to deal with data (e.g. cultural heritage data, scientific data) that are syntactically and semantically heterogeneous, multilingual, multicultural, semantically rich, distributed and highly interlinked. Making this content mutually interoperable so that it can be searched, accessed and reused in the long time is a big challenge for DP involving different levels of interoperability. On a syntactic level, it is needed to harmonize different character sets, data formats, identification syntaxes, notations and collection records adopted in different collections but also to agree on communication protocols for information exchange between content providers. At the level of semantic interoperability, different metadata standards are in use by different institutions to describe the same type of content, metadata formats may be interpreted differently, data is encoded at different levels of precision, vocabularies and ontologies used in describing the content are different and ontology alignment and mapping is hard to completely automate. The multi-organizational and multidisciplinary nature in which content is collected, maintained and published poses new issue of organizational interoperability for DP dealing non only with formats and technical standards but also with different policies, rights and restrictions management, mandates, roles and responsibilities. Interoperability appears a complex and multi-layered concept and a crosscutting concern [9], which encompasses a multidimensional spectrum of aspects ranging from more technological aspects to include several dimensions of the digital preservation universe (e.g. users, policies, legal issues, disciplines). Moreover, different communities and disciplines may have very heterogeneous interoperability requirements since their needs with regard to data management and curation vary considerably. It follows that devising an appropriate solution to the digital preservation interoperability challenges is far from being a merely technical problem and the diversity of the community requirements makes it impossible to aim for a single strategy or system for economical, political, organizational and disciplinary reasons. Interoperability is crucial to address issues like access, provenance, citability, data quality assessment and many others, going far beyond the technical level to embrace a much wider horizon where organizational, social and business strategies must be taken into account in considering effective solutions. If an all-encompassing perspective is taken, including technical, social, organizational and many other factors, a comprehensive picture of this complex landscape can be provided, enhancing the understanding of its faces and orienting strategies for finding specific solutions.

[2]available at `http://www.alliancepermanentaccess.org/wp-content/uploads/downloads/2013/03/APARSEN-REP-D25_1-01-1_7.pdf`

[3]`http://cordis.europa.eu/fp7/ict/telearn-digicult/report-research-digital-preservation_en.pdf`

3. INTEROPERABILITY INITIATIVES

As a first step of our diagnosis of the ecosystem of interoperability initiatives and solutions in DP, we performed an analysis of ongoing and past projects and initiatives covering interoperability issues related to (or relevant for) digital preservation. The aim of the investigation was to produce a database of projects and initiatives to be made publicly accessible within the APARSEN NoE and maintained updated in the long term as a collaborative tool to raise awareness and understanding within the DP community. We collected information about 64 projects and initiatives, clustered around eight macro-areas:

1. **Digital Preservation Conceptual Models and Interoperability Frameworks**: in this category we included the main digital preservation projects, which addressed interoperability issues by defining shared conceptual models or developing interoperability framework architectures. This group contains 1) early research projects in the field of DP (e.g. DELOS) focused on the definition of basic concepts and shared conceptual models as fundamental ways to enable interoperability of the various content holders (mainly digital libraries and archives) and rising awareness about the theoretic basis for the key preservation concepts and entities, 2) later-stage projects which addressed interoperability by developing solutions to integrate digital preservation modules into framework architectures to enable the interoperation with other systems. Examples of this kind of architectures are the PLANETS Interoperability Framework for preservation actions, the CASPAR Integrated Framework based on the OAIS reference model, and the integrated preservation framework using grid-technologies of SHAMAN.

2. **Data Infrastructures for E-Science**: E-science infrastructures represent a key strategic area for digital preservation and a rich source of interoperability challenges. First of all, they are of crucial importance to significantly enhance science in many areas, promoting research, innovation and enabling new ways of collaboration and resource sharing. However, the realization of the innovation potential of these infrastructures, strongly depends on the creation of an interoperable data sharing, re-use and preservation layer. Secondly, these infrastructures may represent robust components to support digital preservation services for science data in general (see the PARSE.Insight project) or in specific domains (see for example the SCIDIP-ES project in the earth science domain). This macro-area clusters existing initiatives that aim to promote interoperability in specific e-science domains through the implementation of e-science infrastructures (e.g. INSPIRE, SCIDIP-ES, CLARIN, DASISH) and describe also some relevant initiatives committed to promote and develop reference models and architectures to enable infrastructure interoperability across systems (e.g. iCORDI, EUDAT GEANT, D4Science-II).

3. **Digital Libraries**: In this category, we included some of the most relevant initiatives to address the interoperability challenges in the domain of digital library. Some of these initiatives focused on the development of a common conceptual framework for enabling interoperability between digital libraries (e.g. DL.ORG) or for exchanging specific types of content (e.g. IIIF), others addressed the issue of creating a unique point of entry to distributed content and heterogeneous resources (e.g. EUROPEANA, EUROPEANA GROUP).

4. **Open Repositories**: Open repositories represent another important domain for developing interoperability solutions related to DP purposes. In the recent years, Open Access repositories and their associated services have become an increasingly important component of e-Science Infrastructures. It has been widely recognized that the real potential of open access repositories for e-Science infrastructures lies on the creation of a network of interconnected repositories providing unified access to distributed scientific resources and scholarly content. The creation of this decentralized infrastructures and the development of added-value services on top of it are entirely reliant on interoperability. In this category we included projects and initiatives addressing three main issues: 1) Metadata harvesting and exchange (CRIS/OAR Interoperability Project) ; 2) Infrastructures for digital repositories (DRIVER and DRIVER II); 3) Repository deposit and access (OpenAIRE, Open Access Repository Junction, Open Archives Initiative).

5. **Persistent Identifiers**: Interoperability between persistent identifiers (PIDs) is one of the key challenge for guaranteeing persistent discoverability, accessibility and reuse of digital resources and therefore is of central importance for enabling effective digital preservation solutions [2]. This category includes a remarkable number of initiatives that in the last years focused on persistent identifiers interoperability, for digital objects (PersID, RIDIR, PILIN), for authors (ORCID), for scientific data and related resources (DIGOIDUNA, EPIC) and for entities in general (OKKAM).

6. **Semantic Interoperability and Linked Data**: this category groups some relevant initiatives, which have adopted the Linked data framework to face problems of interoperability related to digital preservation issues in the library context, such as data interoperability, unified data access and interconnecting data silos. It includes library initiatives aiming at exposing their records as Linked Data (LOCAH, CEDAR, LUCERO), promoting the use of Linked data as a Web standard within the library community (W3C Library Linked Data Incubator Group, BIBFRAME) and using semantic web technologies for enabling semantic interoperability of metadata vocabularies (STITCH).

7. **Semantic Access to Earth Sciences resources**: Exploiting the experience of one of the partners of the project (ESA), we included in the analysis also projects and initiatives in the specific domain of Earth Science since its relevance for DP research (see for example SCIDIP-ES project). The analysis focused mainly on the problem of interoperability issues concerning semantic access to Earth Science resources based on ontologies (OTE, OTEG), semantic discovery tools and frameworks (SMAAD), data and metadata sharing (like GEOSS).

8. **Other**: the last category was introduced to include those projects and initiatives which could not fit into one of the previous categories or domains (i.e. EpSOS in the domain of e-Helth, ISA in e-Government)

Each initiative has been described according to the following categories: 1) Name: the name of the initiative or project, 2) Domain: indicates a specific area to which the project or initiative belongs; 3) Timescale: indicates the duration of the project or initiative; 4) Description: provides information about the project or initiative, its objectives and the issues addressed by it; 5) Interoperability objectives: provides a list of the specific interoperability goals addressed by the project or initiative. 6) Link: is the URL of a Website where more information and documents can be found. An overview of the analyzed initiatives is shown in Figure 1. We refer to the project deliverable for more details about each initiative.

4. CHALLENGES

To frame the discussion around interoperability and start to identify interoperability objectives, gaps and recommendations, we collected from partners and other stakeholders a set of interoperability scenarios and challenges. Each scenario has been evaluated (using Likert-type scales) according to three dimensions, i.e. 1) the current situation about the raised issue, 2) the importance/impact of the issue, 3) the level of difficulty to address the problem. Some scenarios have been directly extracted from other deliverables of the APARSEN project (and we will refer to them for more details) and other sources (e.g. the DIGOIDUNA study[4]). The 13 collected scenarios have been organized into the following clusters pertaining different areas of the digital preservation landscape or specific domains (e.g. Earth Science): 1) Persistent Identifiers (PIDs) Interoperability, knowledge discovery and citability; 2) Semantic metadata Interoperability and lifecycle management; 3) Semantic Interoperability in the EO Domain; 4) Provenance and Authenticity Interoperability. For space reason we can not include a full description of all the collected scenarios which are reported in the project deliverable mentioned above. An example scenario about provenance interoperability is illustrated in the following box to give an idea of the adopted approach.

SCENARIO: Exchange and Aggregation of Provenance Information

A sensor e.g. at a satellite, makes some measurements. The measurements are then transferred to a ground station. The data are then processed by a group of researchers, say group A, to produce an image, say img1. The image is then processed by group B to produce a second image, say img2. To produce the complete provenance of the img2 (which may be important for assessing the credibility/authenticity of img2) we have to aggregate the provenance information of each data object and link them appropriately. This aggregation requires having a common model for representing provenance or mappings between the adopted models.

Challenge: Ability to exchange and aggregate provenance information of various processing tasks or transfer/archiving events.

Evaluation: Current situation (bad); importance (high); level of difficulty (fair).

Relevance for DP: Provenance information is of crucial importance for e-Science (e.g. for checking and validating results, for reproducing them, etc). However, even though several solutions for modeling and recording provenance information have been proposed and various mapping between these models have been defined (see for example [11]), their adoption by the various organizations is still scarce. In short, interoperable solutions for enabling exchange and aggregation of provenance information, like methods that can aid the ingestion and management of provenance information, are available but there is a lack of awareness and understanding by the e-Science stakeholder communities of the importance of adopting these solutions. The interoperability issue is more at the organizational and inter-community level than at technical level.

In this section we describe the main interoperability challenges derived from the analysis of the scenarios for each domain of investigation. In the first cluster, called Persistent Identifiers (PIDs) Interoperability, Knowledge Discovery and Citability, the scenarios covered the following aspects: 1) knowledge discovery and data integration through PIDs; 2) author identifiers interoperability; 3) impact and quality assessment; 4) citability of scientific datasets. From these scenarios we derived the set of challenges reported in the following box.

Challenges from scenarios about Persistent Identifiers (PIDs) Interoperability, knowledge discovery and citability:

1. To provide a global resolution mechanism, which ensures that given an identifier of any kind the correspondent resource can be persistently retrieved and accessed. If the resource is not available any more, a matching resource if available (also from a different provider) should be linked.
2. To provide a unique interface to find integrated information across different systems about an identified entity (e.g. a paper) and related entities (related publications, authors, datasets).
3. To create a collection from resources, that belong together (e.g. enhanced publications).
4. To associate multiple identifiers with the same entity (e.g. author) to enable the long term access to the entity or a description of it.
5. To locate all versions of a resource.
6. To find information about authenticity and availability of a resource.
7. To integrate metadata referring to the same resource from multiple sources.
8. To make citation and their relationships more explicit

[4] http://digoiduna.wordpress.com/about/

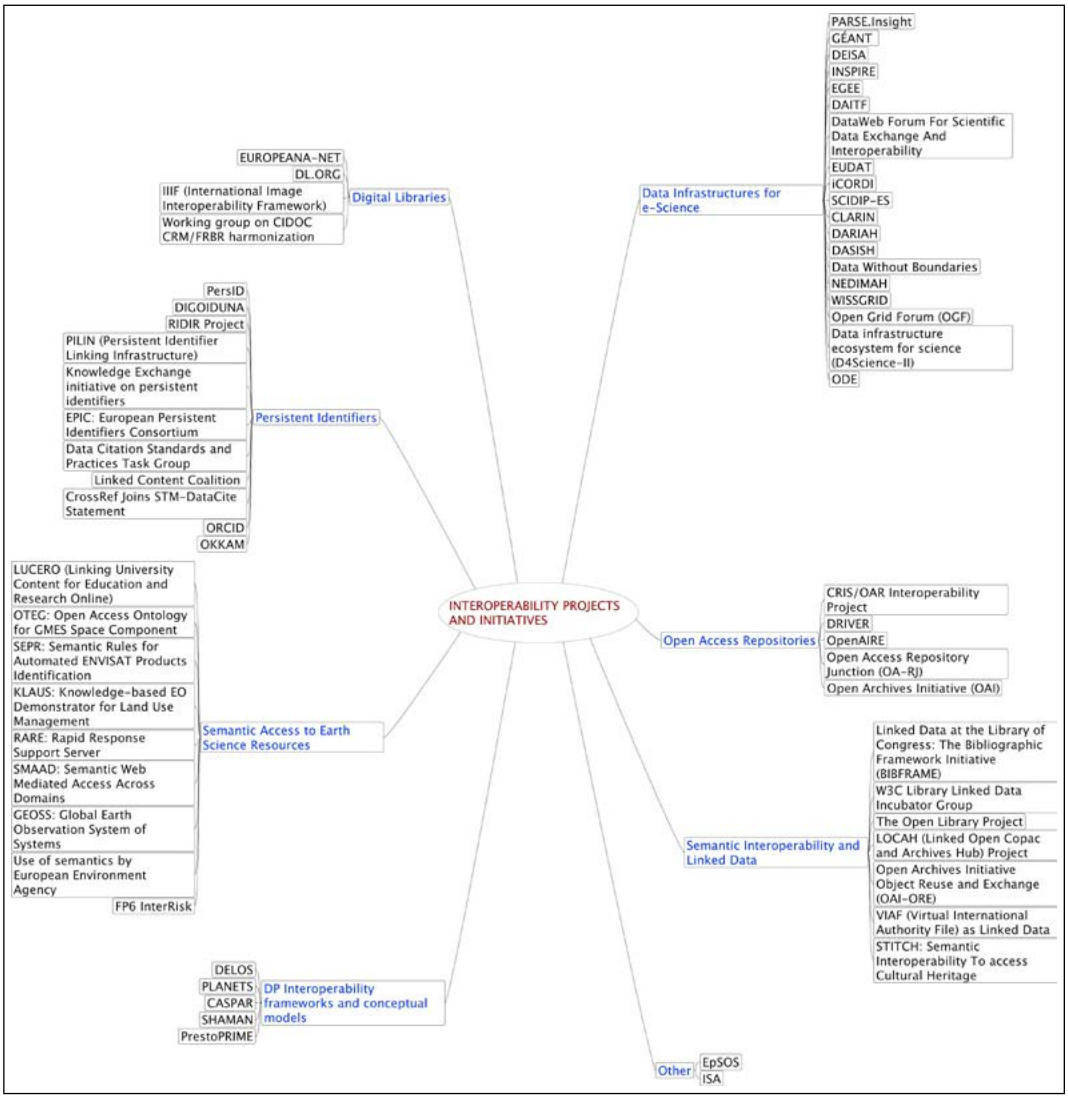


Figure 1: Interoperability projects and initiatives relevant for DP

so that data can be accessed more easily, supporting reuse and verification and strengthen the link between the contributor and data.
9. To define a standard to uniquely identify datasets and manage them as separately citable items.

The second set of challenges pertains the domain named semantic metadata interoperability and lifecycle management. In this domain we included scenarios describing narratives about vocabulary alignment, creation of semantic links between archival collections and other (Web) sources, use of integrated metadata search interfaces across several providers for accessing digitized objects. The following challenges have been derived from the analysis of this second set of scenarios.

Challenges from scenarios about Semantic metadata Interoperability and lifecycle management:

1. To provide mapping between vocabularies, thesauri and categorization systems to facilitate browsing and searching in several library catalogues in parallel with the keywords from any of the used thesauri.
2. Aggregating diverse data sources and performing vocabulary alignment to a common ontology in order to facilitate searching and finding structured results also across multiple languages.
3. To provide metadata mapping between domain-specific metadata models used by different sources.
4. To interlink metadata relevant for digital preservation actions (e.g. metadata about digital objects, their formats, versions, events and agents involved in the events).
5. To aggregate metadata from different data providers

and provide a common way to search for their content using these metadata.
6. To create semantic links between heterogeneous materials from different sources including web resources.
7. To provide identification mechanisms for accessing provenance (metadata) of digital objects and intellectual entities.
8. To develop a common standard for exchange information between institutions adopting different archival systems.
9. To define a framework to relate library publications to datasets that are held by other institutions.

As mentioned above, we investigated the specific domain of Earth Science as an important testbed for DP practices and solutions. From the analysis of the proposed scenarios in this domain, the following challenges emerged:

Challenges from scenarios about Semantic Access to Earth Science resources:

Allowing application domain experts to access the needed EO resources through an interoperable and pluggable architecture, permitting:
1. Data discovery via controlled vocabulary, which would permit the user to search resources through familiar terminology;
2. Direct access to the needed resource, independently where the resources are physically hosted (e.g.: federation of smaller and remote catalogues).

Finally, the analysis of the scenarios provided some interoperability challenges concerning Provenance and Authenticity. In particular the collected scenarios addressed 3 main issues: 1) exchanging and aggregating provenance information of various processing tasks or transfer/archiving events; 2) Querying provenance records of any digital object through services that can fetch and integrate the required provenance information from heterogeneous and distributed sources; 3) finding information about resource authenticity and availability. The analysis of the provenance scenarios produced the following set of challenges.

Challenges from scenarios on Provenance, Authenticity and Rights:

1. To develop a common model for representing provenance information or a mapping solution between different models to aggregate provenance information from different sources.
2. To provide query and retrieving systems and user interfaces to give access to heterogeneous and distributed provenance information.
3. To develop a trusted PIDs infrastructure which guarantees access to authentic digital objects and related provenance information.
4. To define a standard way to expose rights expressions with metadata.

5. SOLUTIONS

The challenges described in Section 4 provide a partial view on the complex ecosystem of interoperability problems in DP. Since the goal of our investigation was to identify the interoperability issues encountered by the APARSEN partners as part of their daily activities and gather the conceptual models, services and standards used by them to address these issues, we deepened the analysis by identifying concrete interoperability barriers, needs and related solutions (i.e. models, standards, frameworks, services) adopted by the partners in relation to the key digital preservation areas investigated within the APARSEN project. The final aim was to describe which are the critical interoperability aspects pertaining a certain area of digital preservation, which main layers of interoperability are mainly involved, which are the interoperability objects that are implicated and finally which concrete solutions (e.g. models, standards) have been adopted to address these issues. The result of the analysis led to define a sort of matrix, which combines different layers of interoperability (e.g. syntactic, semantic, organizational) with the areas of digital preservation (e.g. persistent identifiers, metadata, provenance) and the related interoperability objects and models, providing an interoperability conceptual framework for digital preservation that can be used as a starting point to facilitate practical interoperability solutions and design concrete interoperability services for long-term preservation. To this purpose, we organized the collected information on the basis of a common framework that aims to characterize the problem facets as well as the existing and forthcoming solutions and models. In this way the specific challenges of interoperability within a specific area could be directly linked with the current available solutions, providing a useful working instrument to address concrete issues of interoperability encountered by relevant stakeholders in their daily work activities. The proposed framework includes the following categories: 1) **Digital Preservation area**: indicates the area of digital preservation where interoperability takes place. Examples are preservation services, persistent identifiers, authenticity and provenance. 2) **Interoperability issue/challenge**: a problem of interoperability which hinders a certain task or process in an interoperability context. 3) **Interoperability objects**: are the entities that actually need to be processed in interoperability scenarios. They can include for example the full content of digital resources or mere representations of such resources (i.e. metadata, identifiers). 4) **Adopted solutions/ models/ standards**: are those approaches, which are adopted to address specific interoperability issues/challenges at different levels. An example of a described solution for enabling interoperability for PIDs for authors is shown in Figure 2. The Figure 3 provides a mind map that summarizes the contents of the collected material which has been organized in a matrix containing 58 interoperability solutions. The solutions have been clustered around eight categories identified by colours in the figure: 1) Persistent Identifiers, 2) Provenance, 3) Data Quality 4) Metadata 5) Metadata Harvesting and Information Exchange, 6) Authentication, Authorization, Rights, 7) Preservation Models and Services, 8) Research data deposit, discovery, access, reuse and citation.

6. RECOMMENDATIONS

In order to put theory into practice we have devised four sets of recommendations, which should promote the realization

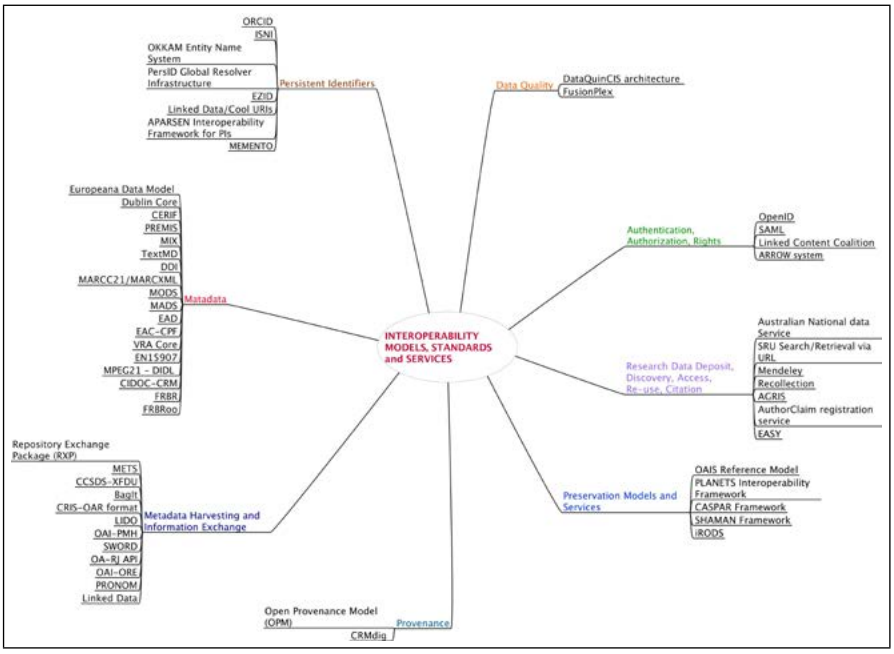


Figure 3: Interoperability solutions for DP

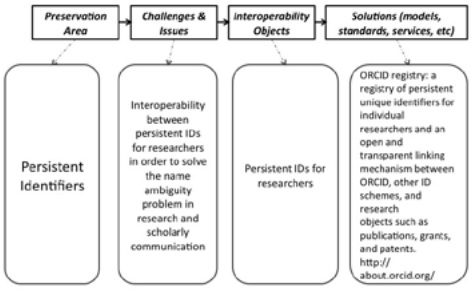


Figure 2: An example extracted from the matrix of solutions.

of interoperability in DP. These are recommendations that are applicable to all the categories of stakeholders and aim at: 1) Fostering the broad adoption of common standards and specifications reducing dependencies, facilitating the interoperation between systems for the entire digital object lifecycle management process and enabling higher-level services on top of standard compliant systems. 2) Promoting the use of appropriate identification systems and their interoperability. 3) Promote the convergence towards agreed common policies and governance models, which foster the adoption of interoperability solutions and trust on them. 4) Ensuring the necessary long-term financial support and the efficient use of economic resources.

6.1 Standards

Standards are considered essential elements of interoperability. The first set of recommendations concerns the definition and adoption of standards as starting point for DP interoperability.

1. Standards are Good: rely on standards in case there are appropriate standards for the digital objects at hand.

The first recommendation about standards states that if compliance to one standard guarantees the achievement of one or more interoperability objectives, then the adoption of the standard is certainly beneficial and recommended. From a dependency point of view, we can say that the standardization essentially makes the dependencies more clear and resolvable.

2. Standards are not a Panacea: be aware that standardization does not vanish the dependencies of the digital objects.

The second recommendation mitigates the first one, specifying that not all the interoperability objectives which can occur in the DP landscape can be addressed through the use of common standards. The obsolescence of a standard, for example, may represent a potential threat for interoperability, e.g. if a standard Y becomes obsolete and there are no longer tools that support it, then a digital object represented through Y could be not reusable any more. The open issue is therefore whether we could tackle the interoperability problem without having to necessarily rely on several and possibly discrepant standards, and whether we can exploit solutions to reduce dependencies and tackle the problems of vanishing or evolving standards. One approach to address

these issues will be briefly discussed in the conclusions to this paper.

3. Define Interoperability Standards through the entire lifecycle of a digital object.

According to the third recommendation, standards should regulate the entire chain of digital preservation steps that form the lifecycle of a digital object from its creation to its re-use through the process of digital preservation.

4. Content providers should adopt standards to ensure that their digital content interoperates with other services and collections allowing the development of a common access point to distributed resources.

The forth recommendation remarks the importance of the use of standards in global information spaces. In these contexts where a huge amount of resources from heterogeneous sources is integrated and made accessible, it is important the adoption of common standards that enable interoperability. Some level of interoperability, for example, is assured at data ingestion by requesting data providers to expose their metadata according to a common standard model for metadata[5].

5. Involve stakeholders in the definitions of standards.

The last recommendation states that since it is difficult to mandate standards, it is easier to work on community accepted standards. Community evolution of standards should be encouraged. A concrete example of a successful coordinated effort between two communities to define a common interoperability standard is the joint effort of the CIDOC Conceptual Reference Model and Functional Requirements for Bibliographic Records international working groups to establish a formal ontology, called FRBRoo[6], intended to solve the problem of semantic interoperability between bibliographic and museum resources, facilitating information integration and exchange.

6.2 Identification

The second set of recommendations deals with two aspects of identification of digital resources: 1) the use of Persistent Identifiers (PIDs) to identify digital resources and other related entities (e.g. authors) and 2) the identity of content.

[5]This approach is used for example by the digital library Europeana (`http://www.europeana.eu/portal/`) which has introduced a cross-domain semantic framework to accommodate the range of metadata standards adopted by the different cultural heritage sectors from which the data are collected.

[6]`http://www.cidoc-crm.org/frbr_inro.html`

Bootstrap an interoperability solution for Persistent Identifiers.

The persistent identification of digital objects (e.g. articles, datasets, images, stream of data) and non-digital objects (namely real-world entities, like authors, institutions but also teams, geographic locations and so on) is becoming a crucial issue for the whole information society and for the development of e-Science infrastructure in particular. However the proliferation of several PIDs systems within different communities and the resulting fragmentation of the PIDs ecosystem has led to an urgent demand for establishing an interoperability solution among the current PID systems to enable the persistent access, reuse and exchange of information across different systems, locations and services. Therefore, actions are needed to bootstrap the convergence towards an interoperability solution for PIDs which open new prospects for advanced value added information integration services. However, since any identifier system is always used within cultural, organizational, geographical and disciplinary boundaries through a technical system, it follows that designing an appropriate solution to the problem of identifiers interoperability involves a number of non-technical issues. This means that any action to bootstrap an interoperability solution needs to work towards systematic implementation of those organizational, political, social and economical factors that foster trust and agreement among the relevant stakeholders.

Elaborate on Information Identity.

Apart from the problem of identifiers, another critical point is the identity of the content. Even though library and archival practice, as well as Digital Preservation, have a long tradition in identifying information objects, the question of their precise identity under change of carrier or migration is still a riddle to science. One theory, developed in the context of APARSEN, that tries to give some light to this aspect is described at [3]. The objective is to provide criteria for the unique identification of some important kinds of information objects, independent from the kind of carrier or specific encoding. The approach is based on the idea that the substance of some kinds of information objects can completely be described in terms of discrete arrangements of finite numbers of known kinds of symbols, such as those implied by style guides for scientific journal submissions.

6.3 Organization, governance and trust

The third set of recommendations concerns the organizational dimension of the interoperability exercise. Since DP is currently conceived as a responsibility to share between different organizations, it has become clear that in such cooperative context, interoperability issues at technical level cannot be solved without promoting an agreement and improving communication at an organizational level.

Raise agreement, increase awareness and social support towards a common interoperability agenda.

Given the complexity of the interoperability exercise in many areas of digital preservation and the variety of stakeholders involved, a common direction must be defined. The involved parties should work together to define a common agenda ensuring a coordinated and interoperable digital preservation ecosystem. The forthcoming VCoE (Virtual Centre of Excellence) of the APARSEN project should play a key role to coordinate the definition of this agenda due to its role in the creation of a common view and understanding about the preservation and interoperability requirements in different preservation domains and research areas. The agenda will define a clear conceptual framework, which will be a pre-requisite for dialogue and achieving consensus across the communities impacted, and serving as the basis for promoting awareness and mobilisation of skills and resources. The common agenda should include at least the following points: 1) Raising awareness about digital preservation interoperability objectives, challenges and available solutions. 2) Promote a cross-boundary view on challenging issues and opportunities. 3) Planning interventions to promote awareness, dissemination and education programs in order to reinforce knowledge and skills on interoperability strategies and solutions.

Foster good interoperability practices.

Spreading good practices for interoperability digital preservation needs to include a more deliberate exchange of lessons learned and case studies documenting the use of emerging solutions, workflows, and techniques across national, organizational and disciplinary boundaries. The analysis and evaluation of scenarios, as well as the identification of prioritized interoperability challenges described in the present document, can be used to benchmark available approaches and systems and identify best practices according to certain identified interoperability objectives. Moreover the use of specific variables of performance (e.g. sustainability of the solution, scalability) can be adopted to develop plans on how to make improvements and adapt specific best practices to specific contexts.

Promote and encourage coordination and collaboration among stakeholder communities around policies and governance addressing interoperability objectives.

The different needs and goals of the stakeholders involved in different areas of digital preservation may hinder the adoption of available interoperability solutions. Therefore, actions are needed to favor the convergence towards common policies and governance, which can help to achieve consensus across the communities. The APARSEN NoE is actively working to promote such a consensus (in particular within the WP35) by defining a methodology for implementing governance structures and data policy management mechanisms to enhance interoperability for permanent access to the records of science.

Work towards global trustable solutions.

Trust is a fundamental issue for DP [7], but it also critical for interoperability solutions working effectively. Actions are needed to promote international agreement on global standards and policies. In this way, users can have evidence of authenticity for world-wide data (e.g. scientific) and resources. The creation of an European Framework for Audit and Certification of Digital Repositories is an example of the actions promoted within APARSEN to build global trust by enabling interoperability between increasingly challenging audit processes in digital preservation.

6.4 Economic

DP poses not only technical, social and organizational interoperability issues but raises also interoperability issues which deals with the economic imperatives of DP which is required to guarantee sustainable results against limited resources. In this section we discuss recommendations which consider the economic aspects of interoperability strategies for DP.

Devise sustainable interoperability solutions.

Securing long-term sustainability of an interoperability solution or service is a key factor for promoting its trust, adoption and success. This can be ensured only if the organization behind it is sustainable and can guarantee the longevity of the solution. This is not simply a matter of finding sufficient funds but concerns many different aspects.

Build a robust community behind the interoperability solution or service.

The first step to establish a sustainable interoperability solution is to gain the support of (possibly) all the involved actors. Interoperability solutions are only possible if cultural heritage institutions, governments, public administrations, research institutions and private organizations work in close cooperation in supporting them, sharing responsibilities and finding adequate business strategies. This strategy has been pursued, for example, by the ORCID initiative (see Section 2.4.4) which worked to gain the support of a broad community including many different stakeholders (like individual researchers, universities, national libraries, commercial research organizations, research funders, publishers, national science agencies, data repositories and international

[7] see http://www.alliancepermanentaccess.org/wp-content/uploads/downloads/2012/09/APARSEN-Trust-Brochure-Low-Res-Web-Version.pdf for a discussion about this topic.

professional societies) before working on devising a technical solution to the problem of interoperability between author identifiers.

Align the interests, roles and responsibilities of the involved stakeholder communities into a sustainable economic strategy and operationalise them in a business model.

The stakeholder participation is also crucial in the definition of sustainable business strategies. To this purpose a business working group including the representatives of all the communities can be created to review membership policies, budget models and investigating funding options to ensure the long term sustainability of the solution.

Provide clear incentives to adopt the interoperability solution.

The lack of clear incentives to adopt a given interoperability solution may threat its use and long-term sustainability. For example, the adoption of shared methods and services by independent organizations may bring costs. Sometimes the costs are financial due to the purchase of hardware or software or for hiring and training staff. In other cases costs are organizational. Introducing a new standard requires inter-related changes to existing systems, altered workflow, changed relationships with suppliers and so on. It is important to make clear the added value of adopting the solution and its beneficial impact in the long-term.

7. CONCLUSIONS AND FUTURE WORK

In this paper we have discussed interoperability challenges and approaches in DP and we have proposed an initial set of recommendations to foster the creation of an interoperable DP ecosystem. The results of this investigation have shown the importance of a coordination among the actors of this ecosystem which goes beyond the technical aspects of implementing a valuable solution, to embrace a much wider horizon including organizational, social, political and economical aspects and implications of adopting it. Raising awareness and increasing a common understanding about the current initiatives and available solutions is a first important step towards this coordinated effort. Therefore, a first future work activity within the APARSEN NoE will be dedicated to make the collected information publicly available, hopefully implementing searching and filtering tool to facilitate the query formulation and navigation of the information space. A second activity will be dedicated to the topic of managing interoperability dependencies. We could say that each interoperability objective/challenge, like those described in the current paper (and deliverable D25.1), is a kind of demand for the performability of a particular task (or tasks). The next step (which will be done in the context of APARSEN) is to identify such tasks, and reflect on their dependencies and on how these can be modelled. The ultimate objective is to propose a modelling approach that enables the desired reasoning, e.g. task performability checking, which in turn could greatly reduce the human effort required for periodically checking or monitoring whether a task on an archived digital object or collection is performable, and consequently whether an interoperability objective is achievable. Such services could also assist preservation planning, especially if converters and emulators can be modeled and exploited by the dependency services. The plan is to follow the general approach described at [6], in particular the approach that supports also modeling converters and emulators described at [12].

Acknowledgements

Work done in the context of NoE *APARSEN* (Alliance Permanent Access to the Records of Science in Europe, FP7, Proj. No 269977). Many thanks to all who have contributed to this report, and in particular to David Giaretta (APA Director, STFC), Simon Lambert (STFC) Veronica Guidetti (ESA), Andrea Della Vecchia (ESA), Emanuele Bellini (FRD), Barbara Sierman (KB), Juha Lehtonen (CSC), René van Horik (DANS).

8. REFERENCES

[1] Tylor A.G. *The Organization of Information*. Westport, CN: Libraries Unlimited, 2004.

[2] P. Bouquet, Bazzanella, R. B. Riestra, and M. Dow. Digoiduna: Digitla object identifiers and unique author identifiers to enable services for data quality assesment, provenance and access. `http://digoiduna.files.wordpress.com/2011/12/digoiduna_final_report_expert_feedback1.pdf`, 2011.

[3] Martin Doerr and Yannis Tzitzikas. Information carriers and identification of information objects: An ontological approach. *arXiv preprint arXiv:1201.0385*, 2012.

[4] Anne Geraci. *IEEE Standard Computer Dictionary: Compilation of IEEE Standard Computer Glossaries*. IEEE Press, Piscataway, NJ, USA, 1991.

[5] Margaret Hedstrom. Exploring the concept of temporal interoperability as a framework for digital preservation.

[6] Y. Marketakis and Y. Tzitzikas. Dependency Management for Digital Preservation using Semantic Web technologies. *International Journal on Digital Libraries*, 10(4), 2009.

[7] Jinsoo Park and Sudha Ram. Information systems interoperability: What lies beneath? *ACM Trans. Inf. Syst.*, 22(4):595–632, October 2004.

[8] Miller Paul. Interoperability: What is it and why should i want it? *ARIADNE Web Magazine for Information Professionals*, 2000.

[9] Jeff Rothenberg. Interoperability as a semantic cross-cutting concern. *Interoperabiliteit: Eerlijk zullen we alles delen*, 2008.

[10] C. Rusbridge, P. Burnhill, S. Ross, P. Buneman, D. Giaretta, L. Lyon, and M. Atkinson. The digital curation centre: a vision for digital curation. In *Proceedings of the 2005 IEEE International Symposium on Mass Storage Systems and Technology*, LGDI '05, pages 31–41, Washington, DC, USA, 2005. IEEE Computer Society.

[11] Christos Strubulis, Yannis Tzitzikas, Martin Doerr, and Giorgos Flouris. Evolution of workflow provenance information in the presence of custom inference rules. In *3rd Intern. Workshop on the Role of Semantic Web in Provenance Management (SWPM 2012), co-located with ESWC*, 2012.

[12] Y. Tzitzikas, Y. Marketakis, and Y. Kargakis. Conversion and Emulation-aware Dependency Reasoning for Curation Services . In *Proceedings of the 9th Annual International Conference on Digital Preservation (iPres2012)*, 2012.

Open Preservation Data: Controlled vocabularies and ontologies for preservation ecosystems

Hannes Kulovits, Michael Kraxner, Markus Plangg, Christoph Becker
Vienna University of Technology
Vienna, Austria
{kulovits,kraxner,plangg,becker}
@ifs.tuwien.ac.at

Sean Bechhofer
University of Manchester
Manchester, UK
sean.bechhofer@manchester.ac.uk

ABSTRACT

The preservation community is busily building systems for repositories, identification and characterisation, analysis and monitoring, planning and other key activities, and increasingly, these systems are linked to collaborate more effectively. While some standard metadata schemes exist that facilitate interoperability, the controlled vocabularies that are actually used are rare and not powerful enough for the requirements of emerging scalable preservation ecosystems. This article outlines key requirements and elements of such an open ecosystem and discusses the starting points for building such a common language. We then present a core set of controlled vocabulary elements for preservation quality, objectives, policies, and components, and demonstrate how these elements are instantiated to connect preservation planning, preservation watch, and experimentation with preservation policies. We show how these vocabularies are used to enable automation and enable the preservation community to collaborate effectively, and point out extension points and future work.

Categories and Subject Descriptors

H.3.1 [**Information Systems**]: Information Storage and Retrieval; H.3.7 [**Information Systems**]: Digital Libraries

Keywords

Digital Preservation, Preservation Planning, Preservation Watch, Linked Data, Ontologies, Semantic Interoperability, Workflows

1. INTRODUCTION

Digital preservation aims at keeping digital information authentic, understandable, and usable over a long period of time and across changing technical environments [20]. In recent years the preservation community has come up with a number of independent systems and tools to solve distinct problems in this domain. These systems include repositories, tools for identification and characterisation, analysis and monitoring, and planning. With the digital preservation domain being strongly community-driven, many of today's available systems have been developed directly by individual or collaborating problem owners.

The capabilities a preservation system needs to possess include planning, operations, and monitoring. Preservation planning focuses on the creation of operational preservation plans that contain a decision for a specific preservation action that fulfills clear objectives. Operations focuses on executing the preservation action on content in the production system along with adequate quality assurance measures. Monitoring focuses on gathering and analysing information from different sources internal and external to the organisation, and checking compliance to the organisation's preservation objectives. This needs to be based on a good understanding of organisational policies, which provide the context for preservation. In general terms, policies guide decisions taken within the organisation to achieve long-term goals.

Preservation decision making is guided by information on specific characteristics of actions and aspects such as file formats. Sources providing this kind of information have been implemented and range from online registries and catalogues for file formats and software, to technology watch reports of recognised organisations. Each information source uses its own way to structure data internally and provide it to users. This variety makes it difficult for preservation systems to truly scale up. Furthermore, the information these registries cover is far from complete and often covers only a specific area.

In recent years, several operational software systems have been presented supporting the discussed capabilities. These systems will often be deployed in conjunction with a repository environment. This requires open interfaces and demonstrated integration patterns in order to be useful in practice. We envisage a preservation ecosystem with the following goals:

1. Connect existing systems in a loosely coupled manner.
2. Enable knowledge discovery and exploitation of potential synergies.
3. Facilitate open growth and community participation.
4. Enable automation and scalability.

Apart from open interfaces and reference implementations, this also requires a common language that provides the necessary semantics for the connecting points of these systems, where they communicate about the same concepts. These include objects, file formats, preservation actions and tools, decision criteria and measures, events and conditions.

In this work, we present a loosely-coupled preservation ecosystem where community members in different roles can use an evolving set of tools to collaborate effectively. These tools are linked on the syntactic and semantic level which enables open growth and eases community participation. This leads us to the following questions, which will be discused here:

1. Which elements in a preservation ecosystem play which role towards achieving information longevity, and what are their information requirements?
2. What are the requirements on a language enabling these elements to be connected in a loosely-coupled manner?
3. How can such a language be leveraged in an evolving ecosystem?

The article is structured as follows. The next section outlines key aspects of preservation systems that require integration and discusses some of the major starting points that provided the backdrop and motivation of this work. Section 3 discusses the SCAPE ecosystem of policy-aware operations, planning, and monitoring components, while Section 4 presents the key elements of the common language that enables these systems to exchange information. Section 5 discusses existing applications and outlines benefits and current gaps. Section 6 summarizes the current state of art and points to future work ahead.

2. BACKGROUND

2.1 Systems and tools

Several different systems with specific aims collaborate in a preservation environment and together support the capability of preserving digital information over time. A plethora of software tools exists that perform identification, characterisation, and migration of digital objects. Characterisation tools such as the Digital Repository Object Identification tool (DROID)[1] and JSTOR/Harvard Object Validation Environment (JHove)[2] perform identification, characterisation and validation of digital objects. The File Information Tool Set (FITS)[3] uses a number of tools including DROID, JHove, and Exiftool[4] and provides a unified output. Examples for migration tools include ImageMagick[5] for converting image files, ffmpeg[6] for audio files, and Ghostscript[7] for converting to PDF. The number of available tools however decreases very fast with increasing complexity of the objects.

The service registry CRiB was one of the earliest attempts to wrap migration tools into web services and making them discoverable and usable [10]. The Planets Testbed strived to provide an experimentation environment to evaluate preservation strategies and sharing results [2]. The SCAPE preservation toolset[8] comprises dozens of migration tools ready to install as Debian packages.

The planning tool *Plato*[9] provides systematic decision making support for preservation planning and implements the method introduced in [5]. It includes a model of relevant aspects, entities, and properties that guide preservation planning and offers a standardised view on decision criteria [12]. An integral part of the preservation plan is the decision for a specific preservation action along with concrete quality assurance. Preservation actions may be entire workflows performing complex operations involving identification, migration, and characterisation tools. The workflow management system *Taverna*[10] allows for the definition, and execution of such workflows on different platforms [13]. The platform *myExperiment*[11] integrates with Taverna and makes it possible to share, discover, and reuse workflows [21]. *Preservation operations* is the activity responsible for the execution of this action and reporting on its success. Preservation plans in Plato are specified following a published XML schema.

The preservation monitoring system *Scout*[12] gathers data from various information sources, analyses it and notifies upon the occurrence of configurable events [9]. Scout is an exensible, evolving knowledge base. The information sources it aims at drawing together include content profiles, format registries, software catalogues, experiments carried out in preservation planning, repository systems, organisational objectives, simulation, and human knowledge [4].

The scalable content profiling tool *c3po*[13] (*Clever, Crafty, Content Profiling of Objects*) analyses the technical properties of large sets of objects based on metadata generated by characterisation tools such as FITS and Apache Tika[14]. The generated profile offers a comprehensive and deep insight into the characteristics of the content set in question. Hence it helps to find outliers, objects with particular properties, and combinations of such. Experimentation in preservation planning aims at using samples from the content set that feature a highest possible coverage of occurring properties. Hence the decision making process directly benefits from a thorough analysis of the content set subject to planning.

Technical registries provide information on relevant aspects such as file formats and risks, software products, potential migration paths, and platforms. Such registries have been available for many years and include: the well-established registry *PRONOM*[15] which is curated by the The National Archives UK, the *Global Digital Format Registry (GDFR)*[16] [1], and the *Unified Digital Format Registry (UDFR)*[17] developed by the University of California Curation Center at the California Digital Library. UDFR is a semantic registry and endeavours to unify the content held by PRONOM and

[1] http://digital-preservation.github.io/droid/
[2] http://jhove.sourceforge.net/
[3] http://code.google.com/p/fits
[4] http://www.sno.phy.queensu.ca/~phil/exiftool
[5] http://imagemagick.org
[6] http://www.ffmpeg.org
[7] http://ghostscript.com
[8] http://github.com/openplanets/scape/tree/master/pc-as
[9] http://www.ifs.tuwien.ac.at/dp/plato
[10] http://www.taverna.org.uk/
[11] http://www.myexperiment.org/
[12] http://github.com/openplanets/scout
[13] http://github.com/openplanets/c3po
[14] http://tika.apache.org/
[15] http://www.nationalarchives.gov.uk/PRONOM/
[16] http://www.gdfr.info
[17] http://www.udfr.org

GDFR. The semantically enhanced *P2* registry [25] pulls together content from PRONOM and enriches it with data from *dbpedia*[18]. The *Conversion Software Registry (CSR)*[19] focuses on software packages that support migration of files. CSR finds migration paths of configurable length based on input and output formats provided by the user.

All these registries have been designed with a specific concern in mind. For example, CSR provides migration pathways with some information on the tools but lacks information about file formats. PRONOM on the other hand gives detailed information about some file formats and selected software tools for migration, but does not provide evidence about their quality. P2 is yet sparsely filled and used in a limited number of scenarios. Its successor, LDS3 (Linked Data Simple Storage Specification)[20], provides an open data publication platform based on Linked Data principles. It does not itself provide a common language for describing published preservation data [24], but of course supports the usage of ontologies.

In reality, the information content of moderated registries tends to be modest in coverage, with many important information needs left unfulfilled. We observe that the design assumption of moderated registries, expecting that a controlled point of reference will be able to cope with evolving facts, leads to knowledge gaps. For instance, a migration tool may be considered as stable in one of the registries, but large-scale experiments conducted by an organisation using the tool on content with specific properties might reveal that the tool crashes in particular cases or does not run on a particular platform. On the other hand, open information models are better positioned to capture the evolving facts and knowledge, and technologies such as RDF provide the opportunities to design an ecosystem made for an open world and evolving technologies.

2.2 Policies

Policies provide the context for successful preservation planning, operation, and monitoring. They govern and control decisions within the organisation. Policies often provide guidance on a high-level, for instance by expressing value propositions to customers. However, there is no clear specification of the exact meaning of a "preservation policy". Sometimes it is used as describing the overall strategy of a cultural heritage institution and its commitment to keep digital material accessible over time. Common examples for policy statements also specify strategies and commitments of an organisation, based on regulatory compliance such as statements in the ISO 16363 Repository Audit and Certification catalogue [14] or on industry practice such as statements collected in a recent preservation policy study [3]. These are well known, but do not separate concerns clearly and often mix objectives with functional means to implement capabilities. Hence, their impact is not always well-understood, and operations based on these are complex to implement.

Most usages of "policies" correspond to what the Object Management Group (OMG) standards call "business policies". According to these standards, policies are "element[s] of governance" that are "not directly enforceable" and they "exist to govern; that is, control, guide, and shape the [s]trategies and [t]actics" [18, 19]. Preservation policies hence should provide the mechanisms to document and communicate about key aspects of relevance, in particular drivers and constraints and the goals and objectives motivated by them. At present, there are no establishd standards for preservation policies relevant to planning or for aspects such as monitoring specifications, Service Level Agreements for preservation operations, or system interfaces. Smith et al. [23] point out that preservation systems operate on a rule level and presents policies that have been translated into rules to be enforced in a repository.

2.3 Standardisation

The digital preservation community has embarked on numerous endeavours towards standardisation of certain aspects required to achieve information longevity. The Planets project[21] presented a conceptual model and vocabulary for representing an organisation's values and constraints[7]. The SHAMAN project[22] has approached digital preservation from an Information Systems point of view and provides a contextualized capability-based view on digital preservation. The SHAMAN Reference Architecture defines the core capabilities Preservation Operation, and Preservation Planning including Monitoring [22]. The SCAPE project is taking this further by implementing appropriate scalable systems that support these capabilities.

A key activity in preservation planning is systematic testing of preservation software. The quality of preservation actions such as tools for migration, but equally of emulators, has to be determined to be able to reach an informed decision for a specific action. The ISO standard 25010 - 'Systems and software engineering - Systems and software Quality Requirements and Evaluation (SQuaRE) - System and software quality models' [16] has its roots in the the earlier ISO 9126 family and defines a hierarchy of high-level quality attributes. It combines characteristics relating to the outcome of interaction when the software product is used in a specific context ("quality in use") and characteristics relating to static properties of software and dynamic properties of the computer system ("product quality") [16]. The ISO 25010 quality model has been adopted in preservation planning to classify decision criteria [12].

Digital objects have certain significant properties that need to be preserved for the objects' performance to be deemed authentic. Significant properties have been extensively analysed in the InSPECT project[23], which has provided a detailed analysis of significant properties of different types of digital objects such as vector images, moving images, and software [11].

Since preservation is a continuous process, a preservation system needs to be capable of monitoring aspects that influence the preservation process. A preservation watch component is designed in [4] for monitoring internal (systems and operations in place, assets and activities) and external (e.g. user communities, technologies, available solutions) influence factors.

The SCAPE project[24] is focusing its work on scalable operations to enable the preservation of large sets of digital information [8]. The components developed in the project use open APIs to enable communication between e.g. plan-

[18] http://dbpedia.org/
[19] http://isda.ncsa.uiuc.edu/NARA/CSR
[20] http://www.lds3.org/
[21] http://www.planets-project.eu
[22] http://shaman-ip.eu
[23] http://www.significantproperties.org.uk/
[24] www.scape-project.eu

ning, watch and repositories. Open APIs make interfaces available to the public and thus enable continuous growth of systems by community participation. Standardisation in this area however needs to go one step further and enable semantic interoperability of components. Information exchanged between these components also needs to be opened up to the community to build synergies, enable knowledge discovery, and move from static to dynamically growing information sources.

2.4 Interoperability

Each of the information sources described above has been developed for particular intended users, types of objects, platforms, and with specific domain and project needs in mind. Hence the way they structure data internally and provide it to users vary. Standard metadata schemes are often adopted to facilitate interoperability of systems. The *Preservation Metadata Implementation Strategies (PREMIS)*[25] working group has produced a technically neutral scheme for preservation matadata. It links intellectual entities, objects, rights, events, and agents to provide a data dictionary. One of the most prominent and commonly used metadata schemes is *Dublin Core (DC)*[26]. DC metadata terms describe resources of various types to enable discovery.

Many systems in the digital preservation domain, including PRONOM and UDFR, adopt linked data techniques to share their data and make them re-usable. At the core of this effort is the Resource Description Framework (RDF)[27]. RDF is a standard model to enable the representation of data and metadata that essentially allows for the expression of subject-predicate-object triples. The Web Ontology Language (OWL)[28] provides further mechanisms for the description of vocabularies or ontologies that define classes and properties. These can be used to annotate, describe and define resources. OWL has a well-defined semantics that facilitates the use of reasoning, supporting ontology management and querying of data. Collections of RDF statements (RDF graphs) can be serialised using a variety of concrete formats including RDF/XML and N3[29], while SPARQL[30] provides a language for querying and manipulating RDF graphs.

3. A PRESERVATION ECOSYSTEM

We observe that many different systems exist that support in digital preservation efforts, and many information sources and tools exist that are directly relevant to the preservation efforts of dedicated systems. Not all of these information sources and tools originate from the digital preservation domain. Components in an open preservation ecosystem need to use standards and appeal beyond digital preservation to enable growth and community participation. They should be built around a simple core instead of aiming for being all-encompassing and overwhelming. It becomes clear that the goal should be to connect and enable rather than impose and restrict. The key domains of the ecosystem in focus are the following.

1. **Organisation.** – The organisation operates an information system, e.g. a *repository*, concerned with the preservation of digital information over time. People acting on behalf of the organisation adopt a number of *tools* in the process of preserving the organisation's digital holdings. These include tools for identification, characterisation, migration, and emulation. The organisation formulates and makes available its *goals and policies* that guide operations.

2. **Solution components.** – This domain includes software tools, platforms, and services addressing real needs of the organisation. These components are developed, maintained, and distributed by commercial or non-commercial solution providers concerned with providing solutions according to market needs. The main building blocks include software tools for identification (e.g. *DROID*[31], and the Linux command *file*), characterisation and validation (e.g. *FITS*), migration (e.g. ImageMagick *convert*), emulation, and quality assurance.

3. **Decision support and control.** – Systems and tools in this domain support the decision making process in preservation planning and exerting control over operations. They are capable of analysing digital objects and providing descriptive information about these objects, monitor changes in the technical environment, and support in the decision making for a specific preservation action. The main building blocks in focus of this paper include *Plato*, *c3po*, and *Scout*.

4. **Community environment.** – Individual people as well as organisations and institutions with a particular concern develop and populate systems that drive the preservation process. These systems contain essential information on aspects relevant to preservation. The main building blocks in this domain include technical registries such as *PRONOM*, but increasingly extend to environments not originally emerging within digital preservation, such as the workflow sharing platform *myExperiment* or public open source software repositories.

Each software system requires information about certain domain entities. For example, *c3po* needs to describe objects it analyses, and *preservation tools* need to report measures. The planning tool *Plato* needs to discover preservation actions, evaluate actions, and describe plans. *Scout* needs to collect measures on all these entitites, detect conditions, and observe events. Finally, decision makers need to describe their goals and objectives in a way understandable by the systems, so that decision support can provide customized advice and support that befits their specific policies and constraints.

4. A COMMON LANGUAGE

4.1 Requirements

From the discussion of the preservation ecosystem and its building blocks it becomes evident that a common language is required to achieve semantic interoperability. The

[25] http://www.loc.gov/standards/premis/
[26] http://dublincore.org/
[27] http://www.w3.org/RDF/
[28] http://www.w3.org/TR/owl2-overview/
[29] http://www.w3.org/TeamSubmission/n3/
[30] http://www.w3.org/TR/sparql11-overview
[31] https://github.com/digital-preservation/droid

expected benefits include the ability to communicate about shared concepts, i.e. query across organisational information, policies, monitoring requests, preservation plans, and preservation components using a single framework. To further align with requirements for preservation systems, such a common language needs to fulfill three key objectives.

1. The vocabulary and instances need to cover elements from different domains and make meaningful connections.
2. The model and its representation need to be accessible to both people and software tools.
3. The model should be based on open standards and Linked Data principles.
4. It should be modular and easily extensible, while scaling freely.

The vocabulary described in this article strives to achieve these objectives by building on a simple core model and applying Linked Data principles. By providing a permanently linked core ontology applying across domains and the ability to extend it continuously, it should provide the appropriate support for an evolving ecosystem. The next sections will describe the core domains of the initial model, while Section 5 shows how the existing models are used across the SCAPE ecosystem to improve information sharing, reasoning and discovery.

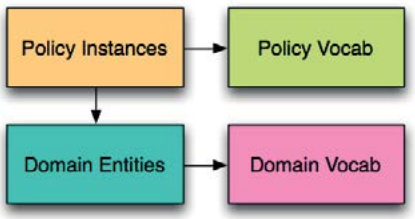


Figure 1: Models

4.2 Control policies

To enable successful communication between decision makers and automated operations, we have developed a core model of specific policy elements that can be represented in a machine-understandable way. We define *control policies* as practicable elements of governance that relate to clearly identified entities in a specified domain model. An element of governance is practicable if it is "sufficiently detailed and precise that a person who knows the element of guidance can apply it effectively and consistently in relevant circumstances to know what behaviour is acceptable or not, or how something is understood"[18]. A control policy contains quantified, precise statements of facts, constraints, objectives and directives about these entities and their properties. Such policies are not directly enforceable. They contain statements that can be fully represented in a machine-understandable model, but the policies are often not directly actionable in that it does not make sense to directly execute them. For example, multiple control statements may contradict each other. A decision making process such as preservation planning translates these policies into a specified set of rules in a plan. This rule set is then actionable and enforceable, and it controls operations. For example, constraints about data formats to be produced by conversion processes can be automatically enforced in a straightforward way.

For expressing control policies, we introduce a *policy vocabulary* that is used to describe concrete control *policy instances*. These policies use vocabulary from a *domain vocabulary* to describe particular *domain entities* such as formats, and content. Figure 1 illustrates these interactions. Figure 2 provides a high-level overview of the policy model including the classes and properties discussed.

Central to a control policy statement is the notion of a preservation case, which links a content set to a user community with particular objectives. Before decision makers embark on a preservation endeavour, the context of "what" has to be achieved for "whom" needs to be established. As Webb et al. describe in [26], an identified set of objects is being preserved for a certain user community, such as images preserved in a library for the general public, or business processes in a company for internal usage to ensure legal compliance. Ultimately, ensuring that the objectives associated with a case are met is the target of preservation planning. To achieve this, objectives need to be associated with measurable outcomes. To this end, we define a "*measure*" as the result of measurement of an "*attribute*". Objectives are thus based on attributes that are represented by measures. Following the definition in ISO/IEC 15939:2002, an attribute is an "*inherent property or characteristic of an entity that can be distinguished quantitatively or qualitatively by human or automated means*" [16, 15]. An example is the attribute *compression* which indicates the compression used. Measures for this attribute include the *compression type* (none, lossless, or lossy), *compression algorithm*, and *compression algorithm covered by patent* which indicates whether licencing fees might occur when using a certain compression algorithm.

In the vocabulary we define a measure as `m(s,r)` with

`s` Scale used for conducting measurement. This includes boolean, number, and ordinal.

`r` Restriction limiting the possible range of measurement values. Specification of a restriction is optional.

Figure 3 shows a set of triples describing a concrete measure for determining the degree of adoption of a certain file format.

We further define a control policy as `cp(m,v,q,mo)` with

`m` A measure pertaining to an authenticity, access, action, or representation instance objective.

`v` A value associated with the measure.

`q` A qualifier (equals, less than, greater than, less or equal, greater or equal).

`mo` A modality that describes whether the particular property-value pair is present or not (must, should, must not, should not).

A sample preservation case is shown in Figure 4. This case relates to a newspaper collection at the Austrian State Archives which is mainly accessed by researchers. It includes an example of a concrete policy statement from this case stating that the degree of adoption of file formats should be ubiquitous.

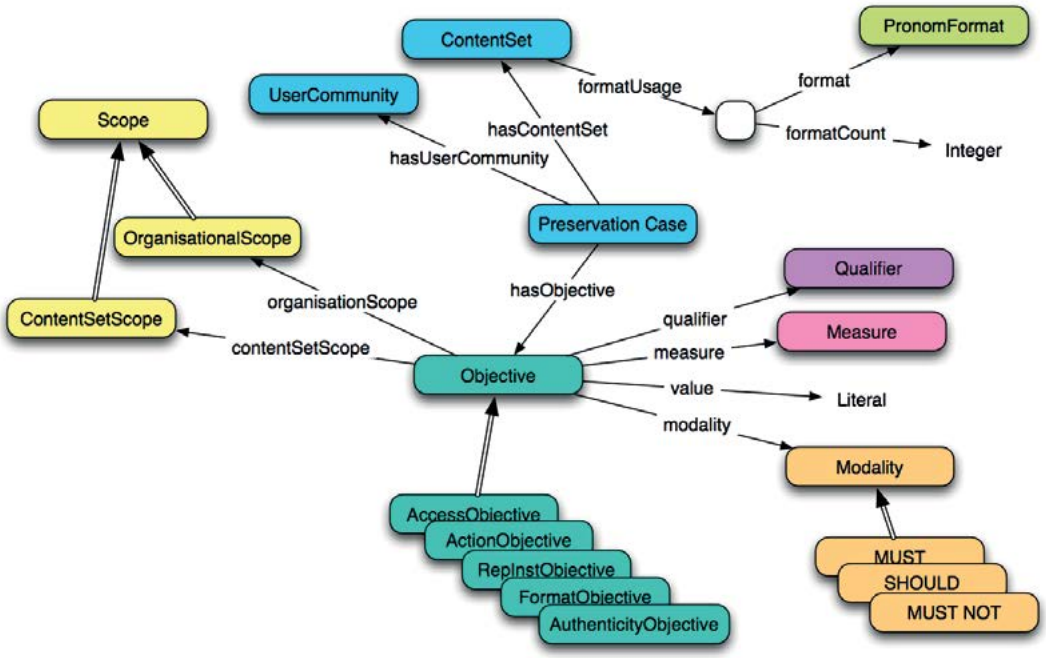


Figure 2: Core model of control policies

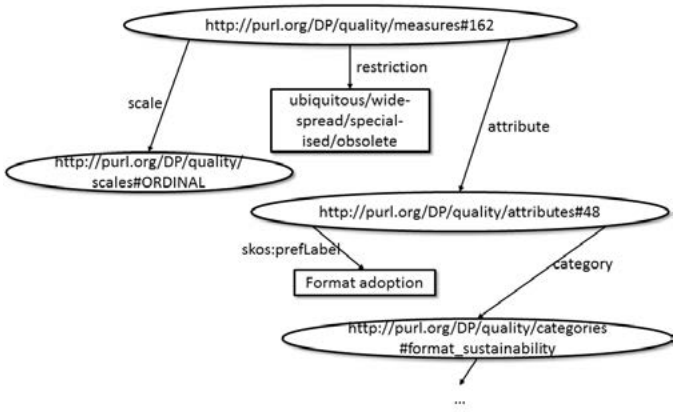


Figure 3: A concrete measure described by triples

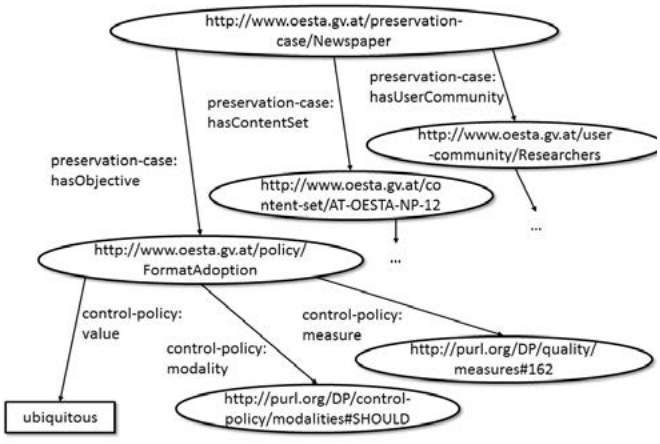


Figure 4: Sample preservation case described by triples

4.3 Domains

4.3.1 Preservation Case

The preservation case documents the particularities resulting from the combination of user community and content set intended for preservation. This includes the time horizon and the goals, objectives and constraints associated with a case. The time horizon will often be determined by legal requirements and contextual issues. To a large extent, access requirements are derived from knowledge on the user community and their used technology. Considering Figure 4 the content set and user community elements provide the extension points to further describe the preservation case.

4.3.2 Objective and Constraint

To be able to preserve the content for a specific user community, clear objectives and constraints on several aspects have to be defined:

- **Format Objective.** This describes an objective referencing a particular property that formats in general should or must have. Most importantly, this corresponds to a risk profile of formats.
- **Authenticity Objective.** This denotes an objective describing the requirements for the preservation of a certain significant property in a preservation case. The set of significant properties can then be used to determine whether a particular preservation action will preserve the authenticity of the performance of each digital object.

- **Representation Instance Objectives** describe objectives referencing a property that representations of content, such as files and bytestreams, should or must have. This includes aspects such as compression, encryption, size, or validity.
- **Access Objectives** This is an objective that describes the requirement for the preservation of a certain characteristic in a particular scenario with respect to accessing the digital object.
- **Action Objectives**, finally, describe constraints on the preservation action process, such as the maximum time or memory resources available or a restriction on allowed licensing.

4.3.3 *Quality*

Preservation cases are associated to objectives, and each objective references a particular aspect of quality in objects, representations, formats, or actions. One of the key activities in preservation planning is the assessment of such quality. Hence, attributes and measures are required to be capable of evaluating preservation solution components and their ability to achieve objectives and minimize risks. Examples include "Format shall be standardised by ISO", and "Image size must be retained".

4.3.4 *Solution*

Software components deployed in the preservation ecosystem require standardised descriptions to enable automation and scalability. For example, planning and monitoring need to discover, evaluate and compose components with minimal manual effort. This will be described in Section 4.4.

4.4 Component profiles

Software tools play a key role in preservation systems. They are deployed for tasks such as format identification, characterisation, migration, or quality assurance. The result of the decision making process in preservation planning is a concrete preservation action to be applied to an identified set of digital objects, including mechanisms for validating the result. Figure 5 shows a high-level view of an executable plan as Taverna workflow with different types of activities (e.g. red circles represent invocation of external tools). Hence, the plan deployed by operations needs to make use of this diverse set of tools and services. Running these tools often requires technical knowledge and expertise in the digital preservation domain. The output generally is (semi-)structured data that neither has a standardised format nor follows a common vocabulary.

To overcome these shortcomings and reduce the overall effort in preservation operations and decision making, an analysis was conducted that identified the following requirements.

1. **Publishing** of components is necessary to allow tool developers and preservation experts to share solution components and expertise needed to create preservation components and enable reuse by others in the community.
2. **Discoverability** of such components is required to enable preservation planning to find the most relevant published components. This allows reuse during planning experiments and in operational plan execution.

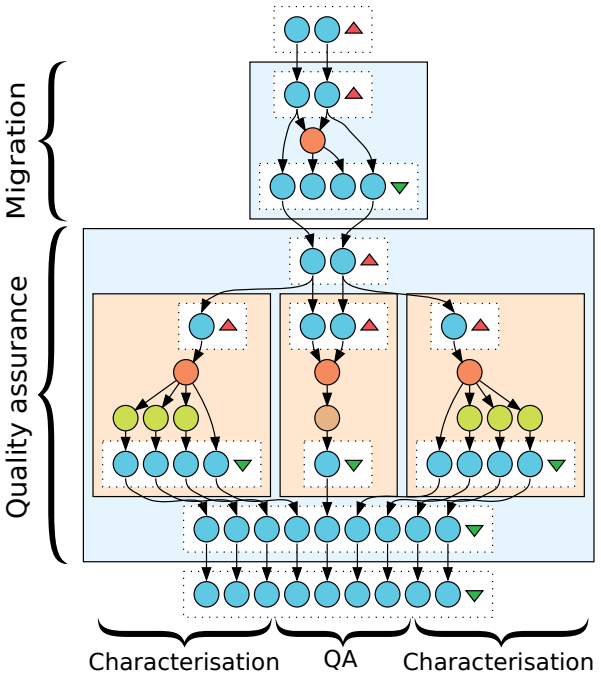


Figure 5: An executable preservation plan

3. **Automated execution** is required to increase the scalability of operations and preservation planning by allowing to run automated planning experiments on representative samples of the content set subject to preservation. This reduces the preservation effort by automating the execution of preservation actions on a large number of tool and parameter combinations. Additionally, it enables automated characterisation and quality assurance of action results.
4. **Reproducibility** is key requirement for trustworthy, evidence-based preservation. Experiments conducted in the course of preservation planning need to be reproducible. Hence, a thorough description of the requirements and dependencies of these tools is required. This is an essential part of the evidence that a preservation plan needs to provide, but equally important for operational deployment.
5. **Standardised output** is required to enable comparability of measures provided by the diverse set of available tools and services. Therefore, the output of components must be well-defined and follow a common vocabulary. This not only allows automated evaluation of experiments in preservation planning, but also enables collecting real world data on tool usage and quality of tool results across organisational boundaries [4].
6. **Composition** is required to allow different components to be combined in an executable plan that can be executed in a repository environment. This should be as easy and automated as possible.

Three main component types of digital preservation tools are in focus:

1. **Migration** components support migrating files to different formats. They must specify supported migration paths.

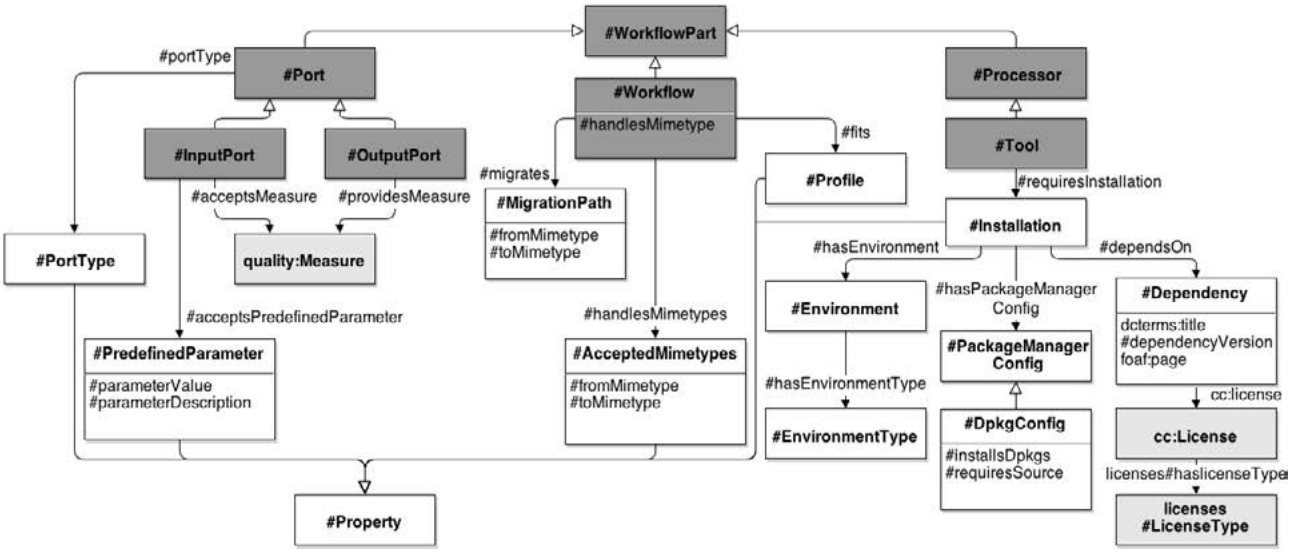


Figure 6: Overview of the ontology http://purl.org/DP/components

2. **Characterisation** components characterise an object and provide specific measures as output. Their specification must contain supported input formats as well as the measures they provide.

3. **Quality assurance** components provide measures that can be used to assess authenticity, validity, cost, and risk. They are split into three subtypes: *Object comparison components* accept two objects as input and provide measures about the degree of similarity. The components need to specify format pairs they support. *Property comparison components* focus on comparing measures and report on the degree of similarity. *Validation components* are used to validate one object against measures. Thus they have to provide supported formats and measures they provide.

In principle, emulation components can easily be added to this ontology. The current focus of workflow development, publication and discovery, however, is on migration and associated characterisation and quality assurance.

Taverna workflows provide a common, platform independent language to execute command line tools and other services and perform pre- and postprocessing on data. All components have to specify the environment they require as well as dependencies needed to execute. Taverna workflows and contained workflow parts can be annotated with human-readable free-text annotations[32]. More complex metadata can be added as semantic annotations based on RDF. The workflow sharing environment myExperiment[33] is an established platform for publishing and discovering workflows and supports querying by annotations.

Preservation components are built on top of Taverna workflows. The Taverna Workbench supports creating components according to component profiles and publishing them to a component catalogue. It also provides basic validation against profiles. Component profiles allow to define the interface and required metadata of workflows as XML documents[34]. As part of the metadata specification, they also define the ontologies used for semantic annotations.

For preservation components, the new ontology `http://purl.org/DP/components` provides a vocabulary to annotate workflows with necessary metadata. Figure 6 shows its classes and properties. The ontology contains classes for the workflow parts that can be annotated. The ports, the workflow itself and associated processors, in the common case preservation tools, each have properties that link them to annotations. For example, a workflow fits a specific profile (such as migration), hence supports a certain set of migration paths, and handles specific mimetypes. Input and output ports are linked to measures in the quality ontology. Annotations can either be literals, individuals already defined in the ontoloy, or more complex RDF graphs from the ontology.

All components must be annotated with the profile they fit. If external tools are used in the component, it must provide the metadata needed to enable execution. These dependencies are modeled as *Installations*. Installations can be used in an *environment* and describe their dependencies, including the license. Further configuration for package managers can be provided to allow automated installation of the dependencies.

5. SUMMARY AND APPLICATIONS

The last section introduced a controlled vocabulary for preservation cases and associated objectives, quality, and solution components. This common language enables interoperability between the building blocks of the preservation ecosystem. A pictorial view of the ecosystem and its building blocks is shown in Figure 7. These include the software systems *Plato*, *Scout*, *c3po*, and *myExperiment* platform which are key elements of SCAPE. The policy vocabulary we proposed is the connecting element between these software systems. The organisation specifies control policies for a spe-

[32] `http://dev.mygrid.org.uk/wiki/display/taverna/Annotations`

[33] `http://www.myexperiment.org/`

[34] `http://ns.taverna.org.uk/2012/component/profile/ComponentProfile.xsd`

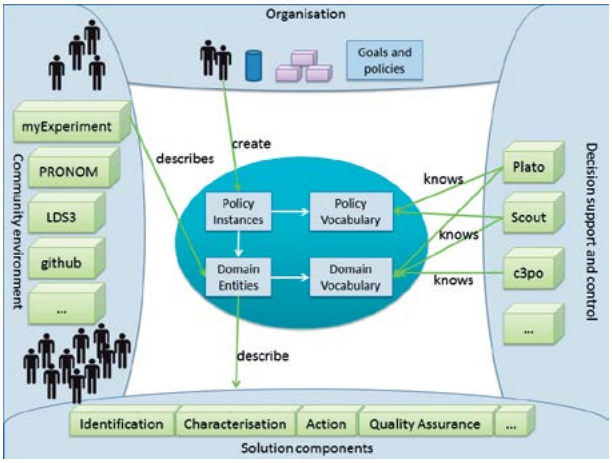


Figure 7: The SCAPE Preservation Ecosystem

cific preservation case, i.e. an identified set of objects that is intended to be preserved for a specified user community. This enables Scout to detect violations and trigger Plato to create a preservation plan for the identified content set. Plato implements the policy vocabulary and relevant domain vocabularies. Hence, Plato can directly incorporate the organisation's objectives, constraints, and directives into planning. c3po creates content profiles using elements from relevant domain vocabularies to describe the digital objects in the content set. This content profile constitutes an essential part of the plan. The component profile allows Plato to query for relevant components on the platform myExperiment, but it also allows standardized specification of executable preservation workflows and their deployment onto target repository environments.To illustrate how the common language is used to improve the effectiveness and efficiency of planning, monitoring and operations, this section discussess several use cases in turn.

Creating policies. To specify control policies, the decision maker leverages the existing vocabulary of domain and policy constructs. Most common policies on this contextual level are until now implicit in the organisational context and used to be discovered in tedious activities within preservation planning [6, 17]. Making these goals and constraints explicit on a higher level with standardised vocabulary enables the decision support tools to offer much more effective support. Tool support for the formulation of policy statements is currently being developed to guide decision makers through a progression of statements that comprise a preservation case. These policies can then be stored in the planning component Plato and the monitoring component Scout, both of which are making use of this organisational context in specific ways:

Automated detection of policy violations. Scout is able to correlate statements in a preservation policy model with the information obtained about the state of affairs in a repository. This most importantly includes the content profile created by *c3po*, which can be queried for violations of specific objectives. For example, the existence of encrypted or compressed files may be not desired. Detecting the existence of such a mismatch causes a notification event to be raised to the attention of the responsible decision maker.

Objective tree construction for evaluation. Upon detection of a non-conforming state or a risk, a mitigation strategy can consist of developing a preservation plan using *Plato*. This in turn is greatly eased by the policy awareness of *Plato 4*, which is able to derive the entire tree of objectives, and measures used for evaluating alternative actions from the control policy model.

Discovery of action components in Plato 4 is enabled through the myExperiment site, where applicable components can be queried, downloaded and executed in a test environment, using the dependency specification to automate installation of required packages. Experimental information that is gathered about the behaviour of tools in real environments on the actual data is associated to the well-defined measurement ontology, which enables cross-linking of cases within an organisation, but also across organisations. Aggregate statistics will in the future be published and can be monitored in Scout, which in turn will enable proactive recommendation of likely successful candidates based on the policies the decision maker's organisation.

For a thorough evaluation of improvements achievable by the integration of the policy vocabulary into Plato 4, we want to refer to a recent controlled case study we carried out with the State and University Library Denmark [17].

6. CONCLUSIONS AND OUTLOOK

This article discussed the information requirements of key building blocks in a preservation ecosystems and showed how controlled vocabularies and ontologies can be leveraged to connect these systems in a loosely-coupled manner to improve knowledge discovery and automation.

We outlined the key systems Plato, c3po, Scout, myExperiment, and Taverna and introduced a common language as connecting element. To enable a loosely-coupled preservation ecosystem where the preservation community can use continuously maturing software tools and collaborate efficiently and effectively, the common language facilitates the systems to be linked on the syntactic and semantic level. We introduced a policy vocabulary based on open standards including RDF and OWL, which enables the ecosystem building blocks to be linked. Concrete policy instances expressed using the policy vocabulary link entities from other domains. Scout can detect policy violations and trigger planning for a specific content set. Decision makers act upon this notification and create a preservation plan.

The current vocabulary presents an important milestone. Current work is geared towards linking in additional existing ontologies to include aspects such as software properties covered in the the Software Ontology (SWO)[35]. On the other hand, we are developing higher level ontology concepts closely linked to preservation intent statements [26]. This aims at dramatically reducing the level of detail required to define objectives related to preservation cases. For example significant properties making up the "appearance" of digital documents can be identified and grouped. An ontology pulling together these properties could reduce the effort of curators to defining "*Appearance must be preserved*" instead of having to deal with the individual technical properties. The decision support system can then derive the set of measures required to assess the authenticity with respect to appearance of specific documents.

Finally, current implementation work on Plato and Scout

[35] `http://theswo.sourceforge.net`

is focused on leveraging this language further.

- Publication of quality assurance components using annotations that specify standardised measures enables Plato to integrate automated evaluation in the experiment workflows and include service-level agreement (SLA) specifications in the generated preservation plans.
- The execution of these generated plans can then be monitored for compliance to the SLAs specifications expressed using the domain vocabulary.
- Additionally, experience sharing on public data endpoints will enable the monitoring of risks and opportunities connected to components and quality measures.

Acknowledgements

Part of this work was supported by the European Union in the 7th Framework Program, IST, through the SCAPE project, Contract 270137.

7. REFERENCES

[1] S. L. Abrams. Establishing a global digital format registry. *Library Trends 54 (1) Summer 2005*, pages 125–143, 2005.

[2] B. Aitken, P. Helwig, A. N. Jackson, A. Lindley, E. Nicchiarelli, and S. Ross. The planets testbed: Science for digital preservation. *Code4Lib*, 1(5), June 2008. See http://journal.code4lib.org/articles/83.

[3] N. Beagrie, N. Semple, P. Williams, and R. Wright. *Digital Preservation Policies Study: Final Report*. HEFCE, October 2008.

[4] C. Becker, K. Duretec, P. Petrov, L. Faria, M. Ferreira, and J. C. Ramalho. Preservation watch: What to monitor and how. In *9th International Conference on Preservation of Digital Objects (IPRES 2012)*, 2012.

[5] C. Becker, H. Kulovits, M. Guttenbrunner, S. Strodl, A. Rauber, and H. Hofman. Systematic planning for digital preservation: Evaluating potential strategies and building preservation plans. *International Journal on Digital Libraries (IJDL)*, December 2009. `http://dx.doi.org/10.1007/s00799-009-0057-1`.

[6] C. Becker and A. Rauber. Preservation decisions: Terms and Conditions Apply. Challenges, Misperceptions and Lessons Learned in Preservation Planning. In *Proc. JCDL 2011*, June 2011.

[7] A. Dappert and A. Farquhar. Modeling organizational preservation goals to guide digital preservation. In *The Fifth International Conference on Preservation of Digital Objects (iPRES 2008)*, 2008.

[8] O. Edelstein, M. Factor, R. King, T. Risse, E. Salant, and P. Taylor. Evolving domains, problems and solutions for long term digital preservation. In *Proc. of iPRES 2011*, 2011.

[9] L. Faria, C. Becker, P. Petrov, K. Duretec, M. Ferreira, and J. Ramalho. Design and architecture of a novel preservation watch system. In *14th International Conference on Asia-Pacific Digital Libraries (ICADL 2012)*, 2012.

[10] M. Ferreira, A. A. Baptista, and J. C. Ramalho. An intelligent decision support system for digital preservation. *International Journal on Digital Libraries*, 6(4):295–304, July 2007.

[11] S. Grace, G. Knight, and L. Montague. *InSPECT Final Report*. InSPECT (Investigating the Significant Properties of Electronic Content over Time), December 2009. `http://www.significantproperties.org.uk/inspect-finalreport.pdf`.

[12] M. Hamm and C. Becker. Impact assesment of decision criteria in preservation planning. In *Proc. of IPRES 2011*, 2011.

[13] D. Hull, K. Wolstencroft, R. Stevens, C. A. Goble, M. R. Pocock, P. Li, and T. Oinn. Taverna: a tool for building and running workflows of services. *Nucleic Acids Research*, 34(Web-Server-Issue):729–732, 2006.

[14] ISO. *Space data and information transfer systems - Audit and certification of trustworthy digital repositories (ISO/DIS 16363). Standard in development*, 2010.

[15] ISO/IEC. *Software engineering – Software measurement process (ISO/IEC 15939:2002)*. International Standards Organisation, 2002.

[16] ISO/IEC. *Systems and software engineering – Systems and software Quality Requirements and Evaluation (SQuaRE) – System and software quality models (ISO/IEC 25010)*. International Standards Organisation, 2011.

[17] H. Kulovits, C. Becker, and B. Andersen. Scalable preservation decisions: A controlled case study. In *Archiving 2013*. Society for Imaging Science and Technology, 2013.

[18] Object Management Group. *Semantics of Business Vocabulary and Business Rules (SBVR), Version 1.0*. OMG, 2008.

[19] Object Management Group. *Business Motivation Model 1.1*. OMG, May 2010.

[20] J. Rothenberg. Ensuring the longevity of digital documents. *Scientific American*, 272, 1995.

[21] D. D. Roure, C. A. Goble, and R. Stevens. The design and realisation of the $\text{my}_{\text{experiment}}$ virtual research environment for social sharing of workflows. *Future Generation Comp. Syst.*, 25(5):561–567, 2009.

[22] SHAMAN. Shaman reference architecture v3.0. Technical report, SHAMAN project, 2012.

[23] M. Smith and R. W. Moore. Digital archive policies and trusted digital repositories. *International Journal of Digital Curation*, 1(2), 2007.

[24] D. Tarrant and L. Carr. Lds3: applying digital preservation principals to linked data systems. In *9th International Conference on Preservation of Digital Objects (IPRES 2012)*, 2012.

[25] D. Tarrant, S. Hitchcock, and L. Carr. Where the semantic web and web 2.0 meet format risk management: P2 registry. *The International Journal of Digital Curation*, 1(6):165–182, June 2011.

[26] C. Webb, D. Pearson, and P. Koerbin. "oh, you wanted us to preserve that?!" statements of preservation intent for the national library of australia's digital collections. *D-Lib Magazine*, 19(1/2), January/February 2013. `http://www.dlib.org/dlib/january13/webb/01webb.html`.

Supporting practical preservation work and making it sustainable with SPRUCE

Paul Wheatley
University of Leeds
Brotherton Library
Woodhouse Lane
+441133435562
p.r.wheatley@leeds.ac.uk

Maureen Pennock
British Library
Boston Spa
Wetherby
+441937546302
m.pennock@bl.uk

ABSTRACT
The SPRUCE Project has applied community oriented approaches to support and sustain digital preservation activity. An emphasis on practitioner requirements and focused agile development has enabled the updating and refinement of key digital preservation toolsets that meet user needs. The capture and sharing of these requirements has provided a detailed snapshot of current curation practice, providing insight into practical practitioner needs for those able to fund and support tool and service development. A variety of collaborative initiatives have developed online resources and forums for supporting digital preservation activity. SPRUCE has begun constructing a toolset to support managers and practitioners in making the case to fund and sustain digital preservation activity.

As SPRUCE enters its final half year, this paper provides an outline of key achievements as well as thoughts on the effectiveness (or otherwise) of some of the more innovative or unconventional approaches taken by the Project.

Keywords
Digital preservation, requirements, agile development, Hackathon, Mashup, business case, collaboration.

1. INTRODUCTION
The SPRUCE Project [1] is a two year collaboration between the University of Leeds, the British Library, the Digital Preservation Coalition, the London School of Economics and the Open Planets Foundation. SPRUCE is funded by Jisc with the aim of supporting digital preservation activity and making that activity sustainable. The project is primarily focused on supporting grass roots preservation activity, by connecting individuals responsible for managing digital data with domain experts, technical experts and a supportive community of peers. Both face to face events and collaboration and support via online communication tools, social networking and web based resources are being employed by the project. Development of a set of resources to support teams and organizations to articulate their case for resourcing digital preservation will help to make this supported preservation activity more sustainable.

This paper focuses in some detail on software tools of relevance to the long term preservation of digital content. There is insufficient space to describe each of these tools in detail, and so it is recommended that readers without experience of these tools utilize a reference resource such as the OPF Tool Registry [2] to provide context to the observations in this paper.

2. SOLVING PRACTITIONER CHALLENGES WITH AGILE DEVELOPMENT AND COLLABORATIVE EVENTS
A core part of meeting SPRUCE's aims has been delivered through the use of agile events, including Mashups and Hackathons. At the time of writing, SPRUCE has delivered 2 Mashups with a third planned for July 2013. The events bring together practitioners (who contribute digital data and preservation challenges) and developers (who apply tools to solve the practitioners' challenges). They support expert attendees in expanding understanding and tackling complex challenges, and help staff from organizations taking their first steps in digital preservation activity. The format of these events was covered in detail in a paper from the AQuA Project at iPRES 2011 [3], so this paper will concentrate on the outputs of these events to date. As well as resulting in many useful outcomes for each individual practitioner or developer, a vital output from the events has been the capture and sharing of practitioner requirements with the wider community.

Practitioner requirements were captured from each SPRUCE Mashup event as well as from AQuA Project Mashups (where the format was first developed), Open Planets Foundation (OPF) Hackathons, and practitioner needs generated by the EU funded SCAPE Project. This totaled over 140 different preservation challenges or "issues", sourced from over 100 practitioners, who represented over 70 different organizations.

Some constraints were placed on the scope and focus of these challenges, mainly related to the scale of challenges that could realistically be addressed in a two or three day event. Practitioners were otherwise left to contribute whatever digital preservation challenges they wanted to have addressed.

All of these challenges (and related descriptions of the data on which they are focused, and the solutions developed to solve them) were captured in different locations on the OPF wiki. SPRUCE collated this data on a single wiki page using Confluence tagging functionality. The result is a detailed record of practitioner requirements and current preservation practice [4] that provides an essential companion to this paper. The solutions to the practitioner derived issues are one of the most obvious and valuable outputs from the SPRUCE Project. Solutions range from fully functioning technical solutions that have since been adopted and embedded in practitioner's organizations, promising prototypes or demonstrators, and also experiments that presented a dead end. For example, a particular tool or approach was explored, but it was decided (often following testing with actual data from the practitioner) that it did not lead to an effective outcome. Capturing the evidence of where a particular tool did not

work well to solve challenges with particular data was seen to be as useful as capturing success stories. Both cases can useful inform (and provide evidence based lessons learned) for other practitioners.

2.1 Understanding and Addressing Practitioner Needs

As the data captured on practitioner needs grew, it was felt that further benefit could be gained from a more detailed understanding of what digital preservation practitioners most needed help with. This became a key focus to explore and report on for the project. What are the priorities for supporting digital preservation practitioners, and what could be done to meet these priorities?

Analysis was performed by SPRUCE on the preservation issues data (i.e. the , with a view to informing the direction of digital preservation tool development. 5 key themes were drawn from the 140+ preservation issues identified by practitioners:

- Quality assurance and repair of damaged or potentially damaged data or metadata
- Appraisal and assessment in order to inform selection, curation and next steps
- Locating preservation worthy data, typically where mixed with other data across shared server space
- Identifying preservation risks in order to inform preservation planning
- A long tail of miscellaneous issues including contextual issues, data capture, embedded objects, and broader issues around value and cost

The overriding focus of these themes is the need to characterize digital data and therefore better understand what it is and what condition it is in. This understanding is typically required before subsequent steps in preservation and curation are undertaken.

Analysis of the practitioner needs provided a review point at which to consider next steps for further exploitation of the best work taken on during the Hackathon and Mashup events, and to consider how the high priority needs could be addressed more effectively. Given the clear need for better characterization it was decided that SPRUCE should host a developer only event which would enable a more concerted effort to update and enhance key digital preservation characterisation tools. Further development work was supported through SPRUCE Awards of up to £5000, which were made available under a funding call for event participants.

A dedicated characterization Hackathon was hosted by SPRUCE and the University of Leeds in March 2013 [5]. It was attended by a group of experts including representatives from many of the high profile, home grown digital preservation characterization tools including: JHOVE, JHOVE2, DROID, FIDO, C3PO and FITS. The theme of the event was to coordinate and combine efforts and technology to improve characterization capability. Four key areas were tackled at the event and are described below.

2.2 Solving the PDF Preservation Problem

PDF issues were a recurring theme in previous Mashup and Hackathon events. The majority of solutions explored the use of Apache Preflight (or related PDFBox libraries), suggesting this technology had considerable potential. The practitioner challenges also highlighted the inadequacy of existing community solutions. JHOVE for example provides very detailed output for PDFs, but without a clear focus on preservation risks (the main practitioner need) and with data on some risks lacking. JHOVE is able to validate a PDF file against the PDF standard. Practitioners wanted to assess a PDF file against an agreed list of genuine preservation risks. Although these two use cases are similar (and indeed overlap) they are not identical; a common misconception which has led to cases of practitioners migrating perfectly renderable PDFs that JHOVE had assessed as invalid (eg. Friese [6]). Therefore the largest of the four groups at the characterization Hackathon wrapped Apache Preflight as a PDF risk analysis tool. An evaluation with large volumes of real data and possible incorporation into key repository technologies to achieve maximum impact for UK Higher and Further Education practitioners (eg. EPrints and DSpace) is being explored as part of the final SPRUCE Mashup, and the OR2013 developer challenge (both in July 2013).

2.3 Consolidating File Format Identification

The “big 3” file format identification tools, DROID, Tika and File, all have their own file format magic which is used to distinguish between each different file format. This leads to the different format identification tools sometimes reporting different results for the same file. Each tool has strengths and weaknesses present in its file format magic. Combining the magic would enable a significant improvement in identification coverage and a reduction in unhelpful and confusing results for the tool users. Addressing this problem would be a big win for practitioners. The group made considerable progress in mapping Tika magic to DROID magic. Although not a complete solution (due to the complexity of the challenge), it provided a large volume of valuable data for the DROID team to collate and enhance the DROID magic, taking us much closer to a single source for file format magic.

2.4 Wrapping Tika for use in FITS and C3PO

The final two groups looked at addressing the complex picture [7] surrounding the key preservation tools: Apache Tika, FITS and C3PO. All of these tools have considerable potential to deliver effective digital collection assessment via automated characterization, but their current status presents a variety of challenges for end users. FITS, for example, wraps a number of out of date tools.

Two groups of developers at the characterization Hackathon focused on incorporating the Apache Tika characterization tool into FITS and C3PO with the aim of making use of the better performance Tika provides and reducing metadata sparsity. Follow up SPRUCE funding awards were granted to address a variety of issues with FITS and C3PO, with the aim of refreshing this toolset. These were ongoing at the time of writing, but considerable progress has already been made (including bringing the wrapped tools within FITS up to date).

The end result should provide a comprehensive assessment and characterization capability with across the board applicability for a large number of practitioners.

2.5 Evaluation

SPRUCE feels it has demonstrated the value of developing software based on comprehensive requirements from practitioners. The real effectiveness of the resulting tool enhancements will become clearer over the final term of the Project, but SPRUCE has clearly demonstrated that significant progress can be made with limited resources if a collaborative and well targeted approach is taken.

The growing popularity and success of activities with some similarity in approach, for example the North American CURATEcamp events [8], reinforces this position. The recent

audio visual focused CURATEcamp day [9] attracted over 150 viewers and a smaller but considerable number engaged via IRC and Google Hangouts.

Home grown preservation tools (meaning those created by this community) are often created with an initial burst of development work, sometimes funded by a specific organization, sometimes with external funding. Whichever the funding source, sustaining the effort, and consequently the tool, can be a challenge. Maintenance and enhancement over time, can however be possible with community contributions and occasional small injections of funding, as SPRUCE has demonstrated.

More effective support in managing tool development, perhaps provided by a coordinating organization, has the potential to make it far more realistic for effective tool maintenance to be performed with these small contributions of effort from across the community (and in particular from occasional Hackathon events). Automated builds, regression testing (essential when making changes and improvements to a complex tool such as FITS) and provision of a consistent test corpora could all play a useful part. SPRUCE partner, the Open Planets Foundation, is seeking to take on this role and has plans to establish supporting activities over the coming months. For example see [10].

3. ONLINE AND REMOTE COLLABORATION

The SPRUCE Project has explored taking some of the positive community experiences from its face to face events and applying them in alternative channels. SPRUCE contributed in a variety of ways to a number of online initiatives. Some were created and launched by SPRUCE, some came about in partnerships with other like minded individuals and organizations, and some were simply promoted by SPRUCE. A single page on the SPRUCE wiki brought together links and publicity to all of the initiatives described below [11].

3.1 Initiatives

A recurring theme at Mashup events, Hackathons and during lively digital preservation discussions on twitter [12] was the need for sharing example files to enable preservation challenges to be collaboratively explored and also to support the development and testing of digital preservation tools. Whilst much larger test corpora, such as the somewhat ubiquitous Govdocs [13], provide material for high volume tool testing, the exchange of small numbers of files exhibiting characteristics of interest seemed to be largely supported via private channels. The OPF established an area on Github as a simple tool to crowd source and manage files of this nature [14]. The only practical constraint is that contributed files must be made available under a CC0 license.

A variety of initiatives relating to Representation Information (RI)[15] were launched during the last year. SPRUCE developed cRIsp in partnership with the UK Web Archive, in order to crowd source RI with as lower barrier to participation as possible [16]. The OPF hosted preservation risk focused pages on its wiki [17]. Jason Scott and the Archive Team launched Just Solve (the file format problem) [18]. And finally, a semantic wiki version of a more formal RI registry was completed by the UDFR project [19]. SPRUCE was not directly engaged with these last three, but it did help to publicise them.

Stack Exchange was quite widely advertised (with support from SPRUCE) as a potentially useful question and answer site for digital preservation topics and via the Libraries and Information Science Stack [20] has accumulated a valuable reference resource for the DP community.

COPTR [21] An ongoing initiative proposed and led by SPRUCE with support from Aligning National Approaches to Digital Preservation is aiming to collate the contents of existing tool registries and reduce some of the unhelpful duplication present in the myriad of existing registries. Four organizations (Open Planets Foundation, Digital Curation Centre (UK), Digital Curation Exchange and Library of Congress /NDSA) who host some of the best existing tool registries have committed to participate in COPTR following production of a wiki based demonstrator [22]. Tool data from these organisation's registries is at the time of writing being collated in advance of production of the COPTR registry.

A single blog post from Barbara Sierman entitled "Where is our Atlas of Digital Damages" [23] prompted two related initiatives. The first captured stories of digital damage, the second focused on images. The latter of these utilized a Flicker group to crowd source images of digital preservation challenges, broken files or "glitch art" [24]. SPRUCE contributed to the latter, publicizing it, collating images from individual contacts and establishing a twitter bot to tweet about new images in the Atlas (which at the time of writing has 132 followers).

3.2 Evaluation and lessons learned

Many of the initiatives were a quiet success, with contributions and interactions from a cross section of individuals from the community. The Format Corpus has gradually received contributions from many quarters (233 commits at the time of writing), and now provides a host of assorted broken files, obsolete files and sets of files exhibiting preservation relevant characteristics (for example the "PDF Cabinet of Horrors" [25]). Contributions of files and usage of files in the corpus was observed during many of the other collaborative events and initiatives described in this paper. Just Solve did not appear to be well supported by the digital preservation community (meaning memory organizations) but delivered the most convincing results of the RI initiatives. cRIsp, launched to an enthusiastic response from the iPRES2012 audience but received a disappointing response from the "crowd". The Atlas of Digital Damages holds 90 images and has 63 members at the time of writing and has received praise in particular as a resource for assisting in communicating the basics of digital preservation visually and in an engaging manner. Although the DP content on Libraries and Information Science Stack was considerable (49 questions) both it, and the proposal for a dedicated digital preservation Stack, were closed after only a short time in beta. Only a quarter of those who signed up to the DP Stack to say they were committing to use the site, actually joined the short lived beta. A poor result, but one that was unfortunately not helped by inflexible moderation and management from Stack itself, that closed the beta without supporting healthy meta discussions with much needed moderation support.

A striking observation for SPRUCE was the substantial lack of formal institutional support for the majority of these initiatives. With a small number of notable exceptions, any success was typically made possible by a cross section of enthusiastic individuals. SPRUCE efforts to enlist support from preservation organizations often fell on deaf ears. When organizational contacts were pushed, it was clear that the unconventional or innovative nature of some of these initiatives was not always viewed favorably. Ownership was also highlighted as an issue.

While organizations were happy to talk the language of collaboration, they were typically reluctant to contribute resources or support to online locations beyond their own organizational URL. This unfortunately explains one of the key reasons behind the current state of online preservation resources where a large number of organizations host very similar information on a variety of topics such as: Getting started in DP, information about DP tools, recommended formats, and so on. As illustrated in the tool registry case (see section 3.1), organizations have not only failed to collaborate in this sphere but they are actively competing with each other. This leaves practitioners struggling to find the support they need. Changing this mindset will be a gradual process requiring direct advocacy and exemplars to illustrate the value of breaking the constraints of walled gardens and competition, and stimulating real collaboration.

Using existing technology and neutral locations to host content related activities was a key theme in the most successful of the initiatives. For example the Atlas utilised Flickr, Just Solve used only a wiki, Format Corpus took advantage of Github functionality. As well as making the setup and management of these initiatives cheap and simple, it provided the community with interfaces and tools with which they were already familiar and were straightforward to use.

4. SUSTAINING THE PRESERVATON ACTIVITY

SPRUCE is building a toolkit of resources that will help managers and practitioners make a convincing case to fund and sustain digital preservation activity. At the time of writing, this toolkit is at an early stage of development, but two ongoing activities are building the evidence base and foundation for this work.

Whilst the main focus of SPRUCE Mashup events has been to understand and solve practical digital preservation challenges, a secondary aim has been to support practitioners in building embryonic business cases. Mashup sessions have included four stages including a benefits brainstorm and alignment exercise, a stakeholder analysis, a skills gap analysis and an elevator pitch. This final stage challenges practitioners to summarize their case in a 60 second pitch to a senior manager. As with the other Mashup activities, results are captured on the SPRUCE wiki [26].

Two SPRUCE funding awards have targeted business case activities, and have taken the form of case studies examining new or expanded digital preservation activity. As well as resulting in sharable exemplar business cases, the process and lessons learnt in their development have been captured. At the time of writing these results are being finalized and will be made available shortly.

5. RECOMMENDATIONS ON CONNECTING THE COMMUNITY

A number of SPRUCE blog posts [27] and presentations have highlighted the challenges of communication and coordination, and what goes wrong when there is inadequate support for these mechanisms that are essential to a healthy community [28]. Duplication and the waste of precious resources are particularly concerning outcomes.

Through its focus on community and collaborative solutions, SPRUCE has made some valuable contributions to the communication required to break away from these negative outcomes. At the lowest level this may simply involve connecting community members with relevant contacts based on an awareness of activity right across the community. For example connecting a user experiencing a particular preservation challenge to an appropriate tool they weren't aware of; making a software developer aware of sources of feedback published elsewhere on some of their code; joining up developers or projects with common aims; heading off new developments, where existing solutions already exist. Connections of these kinds can be important but low key, although they can establish the foundations for far greater partnerships. For example, a weekend twitter conversation between SPRUCE and parties interested in improved format identification led to organic organization of a remote Hackathon, run with members of CURATEcamp [29] that developed new format signatures, facilitated Format Corpus contributions of ebook and video format files and prompted the first step towards opening up the FITS tool to wider community development. The latter of these leading to significant FITS improvements (see section 2.3)

SPRUCE argues that there is a case for a dedicated "digital preservation community manager". SPRUCE has experimented with playing this role and has shown how valuable it is in coordinating activities across the community and in different projects/initiatives. But, as is typical in digital preservation, the role has been funded by a project with a finite lifespan. Ideally this role therefore needs to be adopted by a more sustainable, long term organization such as the OPF, the DPC, or perhaps the ANADP initiative.

6. CONCLUSIONS

SPRUCE activities and related community focused initiatives have met with mixed results so far. Those organizations and individuals that have engaged with SPRUCE activities appear to have got significant value from them. Event feedback in particular was consistently high (for example see feedback responses to SPRUCE Mashups [30]). However a recent SPRUCE Mashup had to be cancelled due to low levels of user registration, suggesting that communication and breaking out to a wider audience remains a significant challenge. Involvement and engagement has not been widespread across the community known to be working in this field.

SPRUCE suggests that barriers to collaboration are gradually being removed and that sufficient value has been obtained from the approaches described in this paper to warrant continued persistence in community collaboration.

7. ACKNOWLEDGMENTS

Thanks to the many people who participated in SPRUCE events and led or contributed to the other collaborative initiatives listed in this paper, without which this work would not have been possible.

8. REFERENCES

[1] The SPRUCE Project, http://wiki.opf-labs.org/display/SPR

[2] OPF Tools Registry, http://wiki.opf-labs.org/display/TR/Digital+Preservation+Tool+Registry

[3] Wheatley, P, Middleton, B, Double, J, Jackson, A and McGuinness, R, People Mashing: Agile digital preservation and the AQuA Project. In: *IPRES 2011: 8th International Conference on Preservation of Digital Objects*, 1-4 November 2011, Singapore. http://eprints.whiterose.ac.uk/43837/

[4] Digital Preservation and Data Curation Requirements and Solutions, SPRUCE wiki, http://wiki.opf-

labs.org/display/REQ/Digital+Preservation+and+Data+Curation+Requirements+and+Solutions

[5] SPRUCE Hackathon: Unified Characterisation http://wiki.opf-labs.org/display/SPR/SPRUCE+Hackathon+Leeds%2C+Unified+Characterisation

[6] Friese, Y, Hunger for Automation – The first migration actions in our Rosetta Digital Archive, IDCC 9, 2013. http://www.dcc.ac.uk/sites/default/files/documents/idcc13posters/Poster213.pdf

[7] To FITS or not to FITS, Petar Petrov, OPF Blog Post, http://www.openplanetsfoundation.org/blogs/2012-07-27-fits-or-not-fits

[8] CURATEcamp, http://curatecamp.org/

[9] AV CURATEcamp day, http://wiki.curatecamp.org/index.php/CURATEcamp_AVpres_2013

[10] Webinar: Software Development with OPF, http://www.openplanetsfoundation.org/events/webinar-software-development-opf

[11] Collaborate with the digital preservation community, http://wiki.opf-labs.org/display/SPR/Collaborate+with+the+digital+preservation+community

[12] Blog summary of twitter discussion regarding obsolete Powerpoint 4 (Mac) files, Rusbridge, Chris, http://unsustainableideas.wordpress.com/2012/10/02/powerpoint-4-0-story-so-far/

[13] Govdocs, http://digitalcorpora.org/corpora/files

[14] OPF Format Corpus, Github, https://github.com/openplanets/format-corpus

[15] OAIS standard, CCSDS, http://public.ccsds.org/publications/archive/650x0m2.pdf

[16] Wheatley, PR, Pennock, M and Jackson, AN (2012) cRIsp: Crowdsourcing Representation Information to Support Preservation. iPres 2012, 1-5 October 2012, University of Toronto, Canada, http://eprints.whiterose.ac.uk/74679/

[17] OPF File Format Risk Registry, http://wiki.opf-labs.org/display/TR/OPF+File+Format+Risk+Registry

[18] Just Solve, http://fileformats.archiveteam.org/

[19] UDFR, http://www.udfr.org/

[20] LIS Stack, http://libraries.stackexchange.com/

[21] COPTR blog post, http://bit.ly/14yVzRz

[22] COPTR demonstrator, http://wiki.opf-labs.org/display/coptr/Home

[23] Digital Damages, Barbara Sierman, http://digitalpreservation.nl/seeds/where-is-our-atlas-of-digital-damages/

[24] Atlas, Flickr, http://www.flickr.com/groups/2121762@N23/

[25] PDF Cabinet of Horrors, contributed to the OPF Format Corpus by Johan van der Knijff, https://github.com/openplanets/format-corpus/tree/master/pdfCabinetOfHorrors

[26] SPRUCE Business Case for DP, http://bit.ly/Z9X8xL

[27] SPRUCE blogs, http://openplanetsfoundation.org/blogs/paul

[28] Bacon, J, The Art of Community, O'Reilly, http://www.artofcommunityonline.org

[29] CURATEcamp file id Hackathon, http://bit.ly/Ye6XQk

[30] SPRUCE Mashup Feedback, http://wiki.opf-labs.org/pages/viewpage.action?pageId=13041673

Creating a Framework for Applying OAIS to Distributed Digital Preservation

Eld Zierau
The Royal Library of Denmark
Søren Kierkegaards Plads 1
1219 København K
ph. +45 33 47 46 90
elzi@kb.dk

Matt Schultz
Educopia Institute
1230 Peachtree Street, Suite 1900
Atlanta, GA 30309
ph. +1 616 566 3204
matt.schultz@metaarchive.org

ABSTRACT
This paper describes work being done towards a *Framework for Applying the Reference Model for an Open Archival Information System (OAIS) to Distributed Digital Preservation (DDP)*. Such a Framework will be helpful for future analyses and/or audits of repositories that are performing digital preservation in highly distributed ways. A great deal of work has already been accomplished toward the Framework itself, including selective community reviews of a white paper, case study interviews with DDP practitioners, and an analysis of OAIS as it relates to DDP. The paper will provide background information on this set of work, describe the research carried out to-date, and explain the proposed Framework components, including concepts and terminology, placement of OAIS functional entities, and roles and responsibilities for carrying out DDP.

Categories and Subject Descriptors
A.1 INTRODUCTORY AND SURVEY

A.2 REFERENCE

E.2 Data Storage Representations: Linked representations, Object representation

E.5 Files: Backup/recovery, Optimization, Organization/structure

H.3.7 Digital Libraries: Collection, Standards, Systems issues

General Terms
Management, Documentation, Design, Reliability, Security, Standardization, Theory.

Keywords
Distributed Digital Preservation, OAIS, vocabulary, functional entities, roles and responsibilities, framework

1. INTRODUCTION
This paper presents work being done towards creating a *Framework for Applying the Reference Model for an Open Archival Information System (OAIS) to Distributed Digital Preservation (DDP)*. Such a Framework will be helpful for future analyses and/or audits of repositories that are performing digital preservation in highly distributed ways.

The use of distribution is actually quite commonplace within the preservation field, but there has not been a commonly accepted definition for "*distributed digital preservation*". As the preservation field has matured, the term "*distributed digital preservation*" has been applied to myriad preservation approaches, spanning initiatives that intentionally build distributed infrastructures as essential components of their preservation models to other initiatives that merely engage in back-up measures for their digital objects or performing all necessary preservation actions within one organization. The reasons for adopting distributed approaches can also be varied in focus and motivation (e.g. achieving cost effectiveness through shared resources [7], expanded support for large amounts of data [6], and general sustainability of the enterprise in the face of contingencies and threats [7]).

In this Framework project, we use the term Distributed Digital Preservation (DDP) to emphasize the practice of applying distribution in intentional ways, both organizationally and technically, for accomplishing digital preservation, for example through geographic distribution, infrastructure heterogeneity, and organizational diversity. A more concise definition, for DDP is: the use of *replication*, *independence*, and *coordination* to address the known threats to digital content through time to ensure their accessibility.

Worldwide, numerous digital preservation initiatives are already engaging highly distributed methodologies, infrastructures, and organizational apparatuses in order to achieve the reliable persistence of digital content. Examples of such organizational and/or technical initiatives include Archivematica (https://www.archivematica.org/), the Danish Bit Repository [5], Chronopolis (http://chronopolis.sdsc.edu/), LuKII (www.lukii.hu-berlin.de/), LOCKSS (http://www.lockss.org/), UC3 Merritt (https://merritt.cdlib.org/), Data-PASS (http://www.data-pass.org/), DuraCloud (http://www.duracloud.org/), MetaArchive Cooperative (http://www.metaarchive.org/), DPN (http://d-p-n.org/), Internet Archive (http://archive.org/), iRODS (http://www.irods.org/), and many others.

These existing DDP approaches, as well as those that have yet to come into being, currently lack common vocabulary and conceptual frameworks for building effective, reliable, and auditable distributed preservation environments. Such agreed-upon terminology and theoretical models would help initiatives to describe and compare their infrastructures and operations. They would also help to increase understanding and awareness of the DDP process both by practitioners and by organizations seeking preservation solutions. Finally, they would provide auditors (including self-auditors) with a crucial foundation for assessing the reliability of a variety of distributed approaches.

The preservation field today relies heavily upon the Reference Model for an Open Archival Information System (OAIS) to provide theoretical and abstract models and vocabulary for digital preservation [3]. This OAIS standard provides a significant portion of the foundation for the Framework this initiative intends to develop. The Framework will elaborate upon the OAIS model to define the methodology and structure of the actions undertaken in organizationally and/or technically distributed preservation

repositories. The Framework does *not* intend to redefine existing standards. Instead it merely seeks to elaborate with additional models and vocabulary.

2. FRAMEWORK BACKGROUND

The awareness of the need for this *Framework for Applying OAIS to DDP* has emerged independently both in North America and in Europe. In North America, discussions first began in early 2010 between MetaArchive Cooperative, Chronopolis and the Library of Congress. Later that year in October 2010, representatives from the Library of Congress and several LOCKSS-based groups met at the 1st Annual Private LOCKSS Network (PLN) conference. It was here that a "constellation group" was formed to discuss the issues and lay the foundation for work related to the Framework.

A formal Working Group was convened in early 2011 that could begin to document the full range of elaborations that may be needed to help apply OAIS to the DDP environment. In this period, the Working Group prepared a Statement of Purpose for the initiative, established a workspace to document DDP use cases and gap analyses[1], outlined a white paper, and solicited participation from a number of DDP practitioners.

Focused conversations began in early 2012 between the Educopia Institute (http://www.educopia.org/) and the Royal Library of Denmark (http://www.kb.dk/) to review the Library's evolved model, known as the IR-BR model [9]. The Royal Library of Denmark is a pioneer in proposing this model, which is an approach to achieving reliable, auditable distributed preservation. The IR-BR model has a great deal of valuable concepts and terminology that present themselves as valuable to the Framework. There is more on the IR-BR model later in this paper.

The Working Group has now grown to include numerous well-known organizations that embody a wide variety of use cases for DDP. These organizations include Archivematica, Chronopolis, Data-PASS, the Danish Bit Repository, DuraCloud, Internet Archive, LOCKSS, MetaArchive Cooperative, and UC3 Merritt

2.1 The Need for a Framework

It is important first and foremost to acknowledge that the Reference Model for an Open Archival Information System (OAIS) does not assume any specific technical or organizational infrastructure, but rather seeks to abstract out the functional and information package requirements that should be achieved in any implementation. It also describes the roles and responsibilities that an archive must undertake, but does not describe where those responsibilities reside, either at an organizational or physical/geographical level.

In practice, some institutions have taken centralized approaches to building an archive that conforms with OAIS - which is to say that their digital objects are ingested and registered into storage resources residing at one geographic and organizational location and stewarded by one organizational center. In this paradigm, the locus of concern, responsibility, and implementation is highly centralized.

Other initiatives have taken distributed approaches to building a repository or network - meaning their digital objects may be ingested and registered into storage resources that reside in multiple geographic locations and that may be stewarded via the use of various distributed services by multiple organizations, all in order to accomplish effective bit preservation. Likewise, a strong case can also be made for the importance and relevance of distribution for the proper hosting, maintenance, and application of services like format identification, validation and migration/normalization guidance (e.g., format registries such as PRONOM, UDFR, etc.). As discussed further in Section 3.3.3. Terminology below, in this paradigm responsibility may be decentralized and span multiple physical and/or institutional locations. Although OAIS is in no way antithetical to such DDP approaches, it does not explicitly define or describe how OAIS principles and models map onto these distributed infrastructures--be they organizational, geographical, or systems-based - each of which protect against different risks and can even be used in combination. A common framework can help to define and make sense of the proper coordinations.

Although OAIS begins to consider issues of interoperability between separate archives (Section 6: Archive Interoperability), it does not explicitly address the range of interactions that may occur between separate organizational entities as part of the work of one distributed digital preservation repository (e.g., collaboratives that share archival management across multiple, distinct institutions (e.g., through the use of embedded peer-to-peer software, distributed micro-services, geographically dispersed Cloud storage services, or other configurations) or even a single organization that manages an archive comprised of distributed infrastructure components—again such as CDL & Archivematica's use of distributed micro-services.

For this reason, early DDP practitioners have encountered the need for additional documentation to describe in greater detail the different functions, roles, and responsibilities that emerge in this distributed landscape. Such a Framework will be helpful for future analyses and/or audits of repositories that are performing digital preservation in highly distributed ways.

2.2 Research Methods

To approach the development of the *Framework for Applying OAIS to DDP*, a number of research activities have and are being undertaken. These include 1) the development of a white paper that was disseminated for peer stakeholder review; 2) a thorough set of case study interviews with several diverse DDP practitioners; 3) a detailed analysis of the Reference Model for an Open Archival Information System with an eye toward bridging gaps in concepts and terminology, proper positioning of functional entities, and roles and responsibilities for DDP; and 4) a review of literature related to DDP and OAIS.

2.2.1 White Paper

The development of the white paper was intended to make a clear case for the need for a Framework for Applying OAIS to DDP. It covered much of what was addressed in the previous Section 2.1 of this paper. It proved to be an extremely useful resource for focusing the proposed work for building the Framework itself and for disseminating information about the Framework amongst numerous peer stakeholder DDP groups and digital preservation experts.[2] Feedback garnered from the white paper review has already played an instrumental role in refining the Framework contents and outline (see Section 3 below).

[1] See http://www.loc.gov/extranet/wiki/osi/ndiip/ndsa/index.php?title=DDP_OAIS_Frameworks (please contact matt.schultz@metaarchive.org to request access to the NDSA wiki).

[2] To request a copy of the Framework White Paper, please contact Matt Schultz or Eld Zierau (mails at top).

2.2.2 Case Study Interviews

Case study interviews were also carried out with each of the Working Group partners, each of whom were asked a series of consistent questions, namely:

- What elements of your organization are distributed? (e.g., management, storage infrastructure, preservation services)
- What has been most challenging about working in these distributed ways?
- If you have used any audit tools (OAIS, TRAC [1], DRAMBORA [4], etc.) have there been any gaps between the concepts and terminology in these tools and the ways you perform your distribution?
- What shortcomings, if any, do you see with these audit tools as you have applied them to your distributed environment?

The responses to these questions were recorded and used to identify distinctive technical and organizational qualities and characteristics of DDP that could be highlighted for the Framework. These elements are also discussed in Section 3 below.

2.2.3 OAIS Analysis

Effectively documenting, within the Framework, the relevant portions of OAIS that have bearing on DDP (and vice versa), is contingent, not only upon our case studies, but upon a thorough analysis of OAIS itself. At the core of OAIS are a set of digital preservation concepts and terminology, functional entities, and roles & responsibilities. The Working Group is reviewing each of these core elements of OAIS and searching for both associations and gaps with respect to DDP as it is best defined both theoretically and through existing case studies.

2.2.4 Review of Literature

Much documentation has been undertaken to describe the proper application of OAIS to digital preservation workflows and repositories. Similarly, much has been written about various distributed implementations for digital preservation. It will be vital that we incorporate relevant information from all such existing publications to ensure proper context, continuity and intelligibility of the Framework for its intended audiences. Users of the Framework will benefit from its association to this broader corpus of information.

3. FRAMEWORK – NOW AND LATER

So far there have been identified a number of topics that are needed in order to make a Framework that can be helpful for future analyses and/or audits of repositories performing distributed digital preservation.This includes *terminology* for DDP and set forth a series of *higher-level concepts, principles* and *guidelines*. Below is a description of the intended audience for the Framework, its scope and publication possibilities, as well as an outline of the Framework's primary proposed components, and the existing and intended contents.

3.1 Intended Audience

There are multiple audiences for this Framework that will need to be kept in view. They span a range of institutional stakeholders responsible for and concerned with the persistence of digital information as well as their designated communities, including governmental agencies, digital libraries and archives, and research data curators, among others.

The primary audience - spanning each of these stakeholders - is those organizations that are seeking to jointly develop or enhance distributed digital preservation (DDP) systems and are in need of guidance on responsible ways of doing so. A second, and equally important group consists of auditing bodies that are seeking to evaluate such DDP systems, and could benefit from the elaborations and interpretations provided by the Framework. Finally, there are the organizations that are seeking to deposit their digital objects in such systems and seeking to understand their operating principles. There may also be other audiences, including those seeking to access and use the digital objects.

3.2 Scope

The Framework currently seeks to address first and foremost the various areas outlined below with respect to both DDP and OAIS. As noted below in Section 4. Discussion and Further Work, focused attention will be given during the later drafting and review stages as to how best to address the relationship of distribution to functional preservation services beyond those of a more generalized repository implementation, which remains the primary focus of the Framework. Such functional preservation services are integral to an overall digital preservation endeavor and encompass things like format registry services and how these may or not be managed and hosted by multiple organizations (e.g. PRONOM, UDFR, etc.). What is likely to be somewhat out of scope (at least initially) is the incorporation of lessons from the general open source software community where collaboration, sustainability, and extensibility (as opposed to hosting and maintaining distributed technologies and administrative resources) are more at issue. The open source community remains important and such lessons are likely to be full of useful insights. Though they are outside of the immediate focus and somewhat broad for the purposes of the immediate Framework, efforts will be made to study what such communities can contribute to the final Framework.

3.3 Publication

This Framework could potentially take numerous forms. One exemplar that already exists is the *Producer-Archive Interface Model Abstract Standard (PAIMAS)* [2], which is a supplemental standard to OAIS itself. Taking this document approach would require review and approval by the Consultative Committee for Space Data Systems (CCSDS) and the International Organization for Standards (ISO). This may not be the most appropriate status for the work as it is currently being proposed. The Framework could perhaps more appropriately exist as a simple community-reviewed document or publication hosted and made available by a respected organization. In addition to such traditional document forms, the Framework could also exist as a modular web resource. There may be other forms. The Working Group will continue to discuss the proper publication forms as the Framework drafting proceeds. Drafting of the Framework is scheduled to be undertaken by the Educopia Institute and the Royal Library of Denmark (financed by the Danish Ministry of Culture) in concert with the Working Group throughout the remainder of 2013, with community reviews scheduled for Fall/Winter 2013.

3.4 Outline

The Framework outline has been developed on the basis of findings from the case studies and reviews, suggestions from the Working Group, stakeholder reviews of the white paper, as well as preliminary analyses of those components of OAIS that are most relevant to DDP (and vice versa). Based on this outline the Framework proposes to cover the following elements.

3.4.1 Introduction

This section will concisely state the purpose and rationale for the Framework and provide an overview of the components.

3.4.2 Background & Overview of DDP

The Framework proposes, first and foremost, to put forth a thorough set of background and overview information on distributed digital preservation (DDP) on its own terms as it has evolved to distinguish itself from more centralized approaches to digital preservation where digital objects are ingested and registered into storage resources residing at one geographic and organizational location and stewarded by one organizational center. This will be helpful for orienting readers to this unique environment, and set the boundaries for effectively applying OAIS for the purposes of the Framework.

3.4.3 Terminology

Many of the case studies have underscored the importance of having improved terminology to describe the parts of a distributed system. Some reviewers voiced support for missing terminology that presents itself as very DDP specific, like some of the proposed supporting terms described below. As mentioned in the introduction, this is more of a task of extending the terminology of OAIS, rather than prescriptively redefining any of the existing OAIS terminology. Nevertheless, where it seems appropriate and helpful, urging new or improved standardization of terminology for the digital preservation community will not be avoided.

So far, the need has emerged for two types of terms to be defined:

- *Major Terms* like the definition of Distributed Digital Preservation, which in its simplest form is the one given in the introduction based on use of *replication*, *independence*, and *coordination* (expanded definitions will also be included in the final Framework)[3].
- *Supporting Terms* which include what is meant by *independence*, and *coordination*, as well as terms like *storage unit*, *storage node*, *storage environment*, *cache*/*pillar*, but also broader definitions of OAIS terms like *replication* and *disaster recovery*.

DDP's definition derives in part from the matured experience of performing proper bit preservation, where bit safety is best ensured by the proper coordination of independent replications of data. The independence of replications requires that replicas be distributed geographically and organizationally and coordinated through timely and effective integrity checking. DDP's definition is also derived in part from the maturing experience of performing functional preservation where sustainable preservation is dependent on shared knowledge and solutions for how and when to do format migration. Here the responsibilities of format registries and the development of trustworthy services (e.g. a migration micro-service) may best rely on their hosting at different organizations. See Section 4. Discussion and Further Work for how this area of distributed functional preservation will be explored throughout the development of the Framework.

A typical example of a supporting term in DDP are any one of those used for *Storage Unit, Storage Node, Storage Environment, Cache, and/or Pillar*. This is the unit that forms the basic storage for a copy of data in a bit repository. This unit is based on specific technology with an organisation around it which can be responsible for basic operations, technology watch, etc. The difficulty here is to arrive at more standardized usage, since most of the terms have other meanings as well, and therefore can be misinterpreted in other contexts.

Replication and *Disaster Recovery* are examples of supporting terms that are extension of OAIS terms. In OAIS the term Replication primarily has a migration context and meaning, whereas in the DDP environment it is more likely to have a proactive integrity measure context and meaning that is closer to OAIS in its treatment of Disaster Recovery. However, disaster recovery as a primary context and meaning for replication also does not do full justice to its usage in the DDP environment. The Framework will make this clearer.

3.4.4 DDP and OAIS Relationships

The Framework can add value for both DDP and OAIS by focusing on the responsible use of distribution at both technical and organizational levels. DDP and OAIS relationships will be related to one another in the Framework by bridging DDP perspectives with OAIS perspectives, especially through analysis of the proper positioning of functional entities, and roles and responsibilities.

- *Functional Entities* will describe scenarios for the placement of the components of an OAIS functional entity across distributed environments, including those that entail the coordination/collaboration of multiple organizations. An example can for instance be found in the IR-BR model (see *Models* below).
- *Roles & Responsibilities* will describe the roles and responsibilities at institutions that are replicating and preserving digital information in geographically and/or organizationally distributed ways.

3.4.5 Case Studies

The purpose of this section of the Framework will be to provide some case study examples of how various efforts are modeled and implemented in distributed ways both organizationally and technically. By no means are these case studies intended to be fully representative of the DDP field at large, nor are they meant to convey mutual exclusivity (i.e., some DDP groups may share case study qualities with others). Nonetheless they will document an impressive array of configurations that can lend insight into the various qualities and characteristics of DDP.

The case studies are based on interviews that were focused towards discovering the aspects of DDP and OAIS as utilized by the various organizations and technical initiatives summarized below.

Archivematica Case Study
Community-Driven Support for Distributed Digital Preservation
Technical collaboration is often integral to supporting a DDP implementation. Archivematica's open source and OAIS-based micro-services infrastructure is highly dependent upon a distributed and collaborative community of both users and developers. This case study will highlight best practices for carrying out coordinated technical approaches to accomplishing digital preservation via such modularized and flexible platforms.

Chronopolis Case Study
Balancing Partnerships for Distributed Digital Preservation
The institutions that comprise the Chronopolis program have structured the administrative workflows necessary to streamline the deployment of resources across three very heterogeneous

[3] We will follow an approach similar to ALCTS (i.e., short, medium, and long definitions). See here: http://www.ala.org/alcts/resources/preserv/defdigpres0408.

organizations. This case study will describe the importance of positioning and coordinating administrative responsibilities effectively across multiple independent organizations that are collaborating toward a shared and distributed repository infrastructure.

Danish Bit Repository Case Study
Shared Flexible Bit Preservation among Institutions for DDP
This case study discusses a platform for shared bit preservation, which allows distribution of copies of data (regardless of whether media is online or offline) as well as services upon them. Shared bit preservation especially challenges roles and responsibilities in connection with how independence among copies of data is maintained, as well as how to ensure that bit preservation solutions for different requirements of bit integrity, confidentiality and availability can be offered.

Data-PASS Case Study
Coordinating Stakeholders for Distributed Digital Preservation
The primary stakeholders in the Data-PASS partnership have balanced institutional independence with the sharing and coordination of resources among institutions participating in collaborative preservation. This case study will explain the importance of cultivating stakeholder buy-in, managing partnership relations, coordination of operations, development of shared practices, and participation in common infrastructure, in order to effectively coordinate actions and ensure sustainability of the overall DDP organizational endeavor.

DuraCloud Case Study
Leveraging Cloud Infrastructure for DDP
This case study will discuss DuraCloud's innovative approach to managing ingest and storage workflows across multiple commercial and public Cloud storage providers. The case study will also discuss the role and importance of external content auditing and making effective use of service agreements in the Cloud service landscape.

Internet Archive Case Study
Fit-to-Purpose Roles & Responsibilities for DDP
This case study will discuss the Internet Archive's impressive efforts to seek international partners that can fill focused roles in the overall set of organizational and technical responsibilities for accomplishing DDP for Internet Archive. Much of this key positioning has been for the sake of optimizing the processes necessary for managing and replicating Internet Archive's large amounts of data.

MetaArchive Case Study
Building Community for Distributed Digital Preservation
Bringing together multiple organizations to accomplish distributed digital preservation is an opportunity to share lessons and develop mutually beneficial technologies. Creating such a community of praxis across multiple organizations can be both challenging and rewarding. This case study of the MetaArchive Cooperative will document the structures and approaches that have made community building possible for this DDP group.

UC3 Merritt Case Study
Dedicated Services for Distributed Digital Preservation:
This case study will explain how the California Digital Library has distributed its UC3 Merritt repository services to optimize the performance of its micro-service oriented architecture and workflows.

3.4.6 Models

The Framework will also investigate different models that can be of help in understanding DDP.

One model being explored can be found in Figure 1, which illustrates the example of distribution for bit preservation [8].

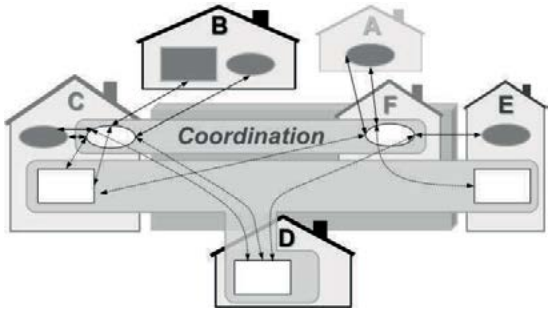


Figure 1. Distribution for Digital Preservation.

The squares could represent a "storage unit" that each hold one copy of data, the houses could represent an organization, and the circles represent services and processes. The dark circles/squares are fully internal, while the white are part of a shared distributed solution. In this case "house" A is only a consumer of a distributed solution, where it does not have any copies of data itself. On the other hand "house" D only has a role as a provider of a "storage unit" to hold a copy of data. Similarly "house F" is only a provider of a coordinated processing service within a distributed solution. This is just to illustrate that DDP can exist in many forms and with varying complexity.

Another example is the IR-BR model shown in Figure 2 [9]. This model was made in connection with a feasibility study for the Danish Bit repository platform.

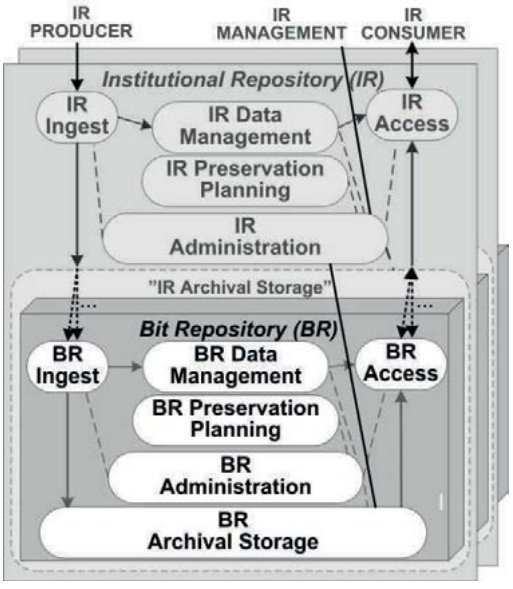


Figure 2. The IR-BR model.

The IR-BR model illustrates a shared Bit Repository (BR) as an OAIS that is shared among a number of Institutional Repositories (IRs) that are also OAIS repositories. Thus the BR is an OAIS within many OAISes.

The idea is that each IR will have their own Technology Watch as part of their *IR Preservation Planning,* which covers ingested file formats. However, the total *Preservation Planning* for the IR and its use of the BR will also cover Technology Watch as part of the *BR Preservation Planning*, thereby covering things like storage media. Furthermore, the BR relies on different "storage units" as depicted in Figure 1. That means the *Preservation Planning* for an

IR covers its own internal *IR Preservation Planning*, as well as the *BR Preservation Planning,* which can be distributed between many organizations with their own technology watches for the media used in their "storage units".

Similarly, additional models may be added to the Framework, in case they can add value in analysis and/or audit of DDP.

3.4.7 Applying to Auditing Methodologies

Many of the case studies address challenges of applying auditing methodologies to their distributed OAIS. Both DRAMBORA and TRAC are inspired by OAIS, thus it is fair to assume that extension and description of the application of DDP to OAIS can be used in addressing some of these challenges. This will include:

- *Trends & Approaches in Auditing* will provide an overview of two dominant approaches to auditing (top-down with TRAC and bottom-up with DRAMBORA) and will highlight some of the trends and directions with the standardizing of the TRAC audit metrics and the rise of audit agencies.
- *TRAC Metrics - How Do They Apply?* will address how areas of TRAC are best interpreted for DDP environments. This description will also address any potential areas unique to DDP that TRAC currently does not address fully.
- *TRAC Auditors - How to Approach DDP?* will address how auditors can prepare to apply TRAC metrics in a DDP environment.
- *DRAMBORA - How to Approach DDP?* will address how DRAMBORA can be best interpreted and used for DDP environments.

3.4.8 Conclusion – Using the Framework

This section will summarize and provide a set of guidelines and suggestions for applying the Framework for both existing DDP practitioners as well as those interested in using such approaches.

4. DISCUSSION AND FURTHER WORK

Though much research has been carried out and the proposed Framework appears well-defined based on contributions from the Working Group and feedback from the broader stakeholder community, there are still relevant questions and topics to be explored as the effort proceeds to the drafting and further review stages. The Framework initiative aims to continue to expand the discussion on this topic and stakeholder participation through subsequent conference papers, poster sessions, and hosted events.

Among the questions and topics that will need further attention, as mentioned above, are the final form(s) the Framework should take and where it should be hosted for broadest uptake and maintenance over time. As also mentioned above, we acknowledge that analyzing distribution for digital preservation in the context of a repository system and environment should be mildly distinguished from collaboration for things like format registries and other open source technology developments. Drawing the proper boundaries is one that likely deserves more attention and discussion.

Further work is also needed to understand the role and needs of consumers of digital objects and trustworthy requirements for access in a distributed digital preservation environment.

Finally, since this Framework initiative does not aim to occupy the place of a standard, the Framework authors are mindful that care will need to be taken to make such relationships clear throughout the resource.

5. CONCLUSION

This paper has explained the beginnings of a *Framework for Applying the Reference Model for an Open Archival Information System (OAIS) to Distributed Digital Preservation (DDP)*.

This is an international initiative comprised of numerous DDP practitioners and stakeholders that have been working steadily and collaboratively since 2011 to clearly define the area of needed work, which includes researching and documenting proper mappings between OAIS and DDP in the areas of *terminology, functional entities,* and *roles & responsibilities*.

This work has resulted in a white paper that has undergone preliminary review by various DDP and digital preservation experts, a series of case study interviews with DDP practitioners, an analysis of OAIS with respect to DDP, and a thorough literature review.

The Framework proposes to include a number of helpful elements, including a Background & Overview of DDP, discussions on Terminology, DDP and OAIS Relationships, a series of Case Studies & Models, as well as a section on Applying the Framework to Auditing Methodologies.

This Framework will be helpful for future analyses and/or audits of repositories that are performing digital preservation in highly distributed ways.

6. REFERENCES

[1] Center for Research Libraries, Online Computer Library Center. 2007. Trustworthy Repositories Audit Criteria: Checklist and Certification v1.0. (also ISO Standard 16363).

[2] Consultative Committee for Space Data Systems. 2009. *Producer-Archive Interface Methodology Abstract Standard,* Magenta Book, May 2004, CCSDS Secretariat.

[3] Consultative Committee for Space Data Systems. 2002. *Reference Model for an Open Archival Information System* (OAIS): ISO 14721:2003.

[4] Digital Curation Centre & DigitalPreservationEurope. 2007. *Digital Repository Audit Method Based on Risk Assessment*, Version 1.0.

[5] Jurik, B. A., Nielsen A. B., Zierau, E. 2012. Flexible Bit Preservation on a National Basis, In *Proceedings of the IS&T Archiving 2012*, Copenhagen, Denmark, pp. 2-7.

[6] Littman, J. 2006. A Technical Approach and Distributed Model for Validation of Digital Objects, In *D-Lib Magazin*e, vol. 12 no. 5.

[7] Rosenthal, D. S. H. 2010. Bit Preservation: A Solved Problem?, In *The International Journal of Digital Curation*, vol. 5, no. 1.

[8] Zierau, E. M. O. 2011. *A Holistic Approach to Bit Preservation*, Doctoral Thesis, Copenhagen University

[9] Zierau, E., Kejser, U.B. 2010. Cross Institutional Cooperation on a Shared Bit Repository, In *Journal of the World Digital Libraries*, vol. 3, no. 1, pp. 11-21.

Realizing the Archivematica vision: delivering a comprehensive and free OAIS implementation

Peter Van Garderen
President, Artefactual Systems, Inc.
202-26 Lorne Mews
New Westminster, BC, Canada
1.604.527.2056
peter@artefactual.com

Courtney C. Mumma
Systems Analyst, Artefactual Systems Inc.
202-26 Lorne Mews
New Westminster, BC, Canada
1.604.527.2056
courtney@artefactual.com

ABSTRACT
Archivematica began in 2008 as a working hypothesis that assumed a comprehensive yet free digital preservation system could be created by matching existing open-source software tools against the OAIS functional model. Five years later the production release of the software is ready to go into production at several major North American archives and libraries, while the beta version is already widely-deployed worldwide. In the absence of a single major funding sponsor, the project management team worked as third-party contractors to several early-implementer institutions that shared the project's architectural and open-source vision while needing to implement an effective and sustainable digital curation solution for the digital content entrusted to their care. From the outset, Archivematica's system requirements were based on an ongoing dialogue within the digital curation community about the gaps between the standards and strategies that were held up as best practice (OAIS, PREMIS, normalization, agile development and so forth) and the ability for the average archivists and librarians to implement them. The iPres conference has proven to be a critical forum for advancing this dialogue and has included papers about the Archivematica micro-services architecture [1] and community-driven development approach [2]. This paper will provide a conclusion to these earlier papers by discussing the key architectural, digital curation and sustainability challenges that the Archivematica project has addressed as it emerged from a working prototype to a full-featured digital preservation system. This includes system scalability, customization, digital repository interfaces, format policy implementation, and a business plan that stays true to the ideals of the free software community.

General Terms
Documentation, Performance, Design, Reliability, Experimentation, Security, Standardization, Theory, Legal Aspects.

Keywords
archivematica, digital preservation, archives, OAIS, migration, formats, PREMIS, METS, agile development, open-source

1. HISTORY
In 2007, the UNESCO Memory of the World Subcommittee on Technology report entitled "Towards an Open Source Repository and Preservation System" concluded that "for simple digital objects, the solution to digital preservation is relatively well understood, and that what is needed are affordable tools, technology and training in using those systems...A practical open source system for digital preservation could, with a little work, be constructed and...this would be of enormous benefit to communities and institutions all over the world." [3] On the report's recommendation, UNESCO offered its support to fund the beginnings of what would become Archivematica, a system that would make it possible and easy to implement in one system what had until then been disparate advances in open source tools for digital preservation. With a trusted digital repository system, memory institutions could preserve the authenticity of their valuable digital records over time.

Nearly concurrently, the City of Vancouver Archives had reached a similar conclusion to that of the UNESCO report. In 2008, the City of Vancouver Archives contracted Artefactual Systems to design and develop a comprehensive digital preservation system that implements the ISO 14721 Open Archival Information System reference model [4]. In 2009, the International Monetary Fund (IMF) Archives also contracted Artefactual Systems to develop a proof-of-concept system based on the work that was being done at the City of Vancouver Archives.

In 2009, the Archivematica project and its partners first translated the OAIS functional model into use case scenarios [5], subsequently developing working prototypes demonstrating implementation of these scenarios. In 2010, Peter Van Garderen introduced Archivematica to the international digital preservation community via his iPres paper, "ARCHIVEMATICA: Using Micro-Services and Open-Source Software to Deliver a Comprehensive Digital Curation Solution" [6] and in a paper for the IS&T Archiving proceedings [7].

As the project advanced, more institutional partners offered to fund features, and what began as tools bundled together into a loose workflow reliant upon Python micro-services and active folders in an operating system became a fluid workflow operated via an elegant, web-based dashboard. Then, in 2012, Van Garderen and Courtney Mumma updated the international

community on the agile, open-source development of the project in their iPres paper "The Community-driven evolution of the Archivematica project" [8]. Now, the Archivematica project is nearing its first production release, which includes features and enhancements motivated and steered by Artefactual's core development team, client partners and users in the community at large. Several clients have successfully deployed beta versions for their individual pilot projects, providing rigorous testing and valuable feedback.

From the outset, Archivematica's system requirements were based on an ongoing dialogue within the digital curation community about the gaps between the standards and strategies that were held up as best practice (OAIS, PREMIS, normalization, agile development and so forth) and the ability for the average archivists and librarians to implement them. The iPres conference has proven to be a critical forum for advancing this dialogue. This paper will provide a conclusion to the earlier iPres papers by discussing how the system has come to fruition. As with each prior release, the first production version of Archivematica will include some new features sponsored by client institutions, enhancements and bug fixes as well as features the Artefactual team considers to be essential for a full-production system. These include system scalability, customization, digital repository interfaces, format policy implementation, and a business plan that stays true to the ideals of the free software community.

2. DIGITAL REPOSITORY INTERFACES

Since the beginning of the digital era, institutions have been building their digital capacity by investing resources and training in content management systems, storage infrastructure, and web components. Recognizing the value of this investment and wishing to bolster rather than replace the systems in existence, Archivematica was conceived as a back-end supplement to manage as-yet unaddressed preservation risks. Since its inception, the intent has been to allow for integration of Archivematica with any number of different access and storage systems. A core design principle is to work with existing collections management tools (e.g. ICA=AtoM, CONTENTdm) and storage architectures (e.g. network storage devices, LOCKSS, cloud storage).

Archivematica is intended to fill the digital preservation services gap for existing repository management applications rather than try to replace them or replicate their functionality. For example, work with our pilot project partners lead to integration of the Archivematica processing pipeline with systems like DSpace, CONTENTdm, Fedora, ICA-AtoM (which is developed in tandem and comes packaged with Archivematica), and Archivists' Toolkit. The partner institutions continue to use the same tools they were already using for collections management, cataloging and public access while Archivematica handles digital preservation services and workflows for the digital materials managed by those other systems.

In one example, Archivematica functions as a "dark archive" for DSpace, providing back-end preservation functionality while DSpace remains the user deposit and access system [9]. For this integration, Archivematica added rules for structuring the DSpace export for ingest, enhancements to the METS file, and an OAI harvesting option [10] that allows for automated ingest of updated descriptions in DSpace. In another example, CONTENTdm integration required changes to the METS structMap including user-supplied structMaps that will allow users to set upload and display order based on logical divisions like book chapters. In addition, Archivematica added a variety of CONTENTdm workflow options for DIP creation and upload [11].

During the development of these and other interfaces, it became clear that institutions' instances of identical systems were unique based on their local configurations. Artefactual's first priority is to integrate with the client-specific configurations, but the community ultimately benefits from a more generic feature that can then be tailored to local specifications. For this reason, Archivematica now includes a generic version of the feature developed alongside the sponsoring client. Moreover, with each integration, Archivematica moved closer to application programming interfaces (APIs) [12] for Ingest, Storage and Access systems. These APIs are the way forward for future integration development.

3. CUSTOMIZATION

With each iteration, Archivematica has been developing methods to make it easier for users to customize their workflows within the constraints of the system. While Archivematica will continue to build and enhance features for customization in future releases, much work has already been done to make the system more flexible.

Users can now change the workflow to skip, automate or include decision points for micro-services, adjust compression algorithm and size for the AIP and pre-select a standard AIP storage location. For instance, the user can decide to skip backing up their transfer or to quarantine the contents for 30 days prior to processing in order to allow for updates to virus definitions in the malware checking tool. If users are ingesting digitized objects, they can set Archivematica to automatically approve normalization or to detect access derivatives included in the SIP, thereby reducing processing time for digital object types that have known behaviour in the system.

Should users have a local tool, proprietary or open-source, that proves better suited within their institution to normalize a particular format to its preservation and/or access copy, they can opt to use a new manual normalization feature either at the beginning or in the middle of the Archivematica workflow. Another option in the normalization workflow is to pick from an ever-expanding set of tools which identify file formats as the basis of normalization actions. Additionally, users can choose to send their transfers to a backlog with enough metadata in the METS file and an accession number so that they can be retrieved from storage to ingest at a later date and even by a different user.

4. FORMAT POLICIES

The format problem is one that, despite the noble efforts of information professionals and hobbyists, does not appear to be solvable in the immediate future. However, advances are happening more rapidly than they were even five years ago, with groups like Open Planets Foundation [13] investing their resources into rigorously testing tools for format identification. Early on, Archivematica was developing media type preservation plans, attempting to discern best practice from the available research at the time. Unfortunately, there was not much information about what institutions were choosing as preservation formats--there still is very little, in fact. It was also the case then and is now that the information out there was in various types of unstructured formats (e.g. webpages, pdf).

Archivematica researchers garnered what they could about significant characteristics and best practices from the varied community of information professionals and from institutional policies. After this analysis, they tested open source tools to implement a two-pronged approach to preservation planning: normalization on ingest and the preservation of the original file to support future strategies such as migration and emulation. Normalization is based on *format policies*, which indicate the actions, tools and settings to apply to a file of a particular file format in order to make a preservation or access copy. The criteria for selecting default formats for normalization in Archivematica are that they must be free of licenses and patent restrictions, have freely available specifications, and be widely used and/or endorsed by major repositories. Preservation formats must also allow for no or lossless compression and there must be open source tools readily available to write and render them.

Archivematica analysts have been closely involved in the digital archives and library community monitoring advances in format identification. Conversations and workshops at events like CURATEcamp [14], national and international conferences and online in blogs and on listservs have highlighted the dearth of certainty about best practices, tools and preservation formats. Because format policies will change as formats and community standards, tools and practices evolve, a Format Policy Registry (FPR) [15] emerged as Archivematica's strategy for the treatment of format policies.

One of the primary goals of the FPR is to aggregate empirical information about institutional format policies to better identify community best practices. The FPR provides a practical, community-based approach to OAIS preservation and access planning, allowing the Archivematica community of users to monitor and evaluate formats policies as they are adopted, adapted and supplemented by real-world practitioners. The FPR APIs are designed to share this information with the Archivematica user base as well with other interested communities and projects.The FPR lists all of Archivematica's default format policy rules and provides valuable online statistics about default format policy adoption and customizations amongst Archivematica users. In the past, the Archivematica project managed all format policy documentation on its public wiki; with the FPR, this information is captured in a structured format (SQL/JSON). Subscription to the FPR (fpr.archivematica.org) via the Archivematica dashboard provides users with notifications about new or updated preservation and access format policies, allowing them to make better decisions about normalization and migration strategies for specific format types within their collections. The FPR will evolve to interface with other online registries (such as PRONOM and UDFR) to monitor and evaluate community-wide best practices. Use of the FPR over time will enrich community understanding of format preservation practices and help to reduce the risk of technology obsolescence and incompatibility.

5. SCALABILITY

Early adopters and testers have consistently asked for scalability metrics; however, without many production betas deployed in client repositories, these metrics were slow in coming. To start, the project team tested distribution of services across multiple processors in order to maximize ingest productivity [16]. While such tests were useful, more rigorous, onsite scalability testing was clearly necessary.

For 1.0, the project staff set up a dedicated testing environment with a full matrix of test parameters [17]. The testing environment begins with several virtual machines set up in a hosted environment, where hardware resources can be scaled up and down between tests. Creating different test configurations allows the project to compile operating system (i.e. cpu and memory usage) and Mysql metrics. With these statistics documented on our public wiki, we can provide metrics to users about scalability and make more informed deployment decisions.

Archivematica has also introduced multiple installation scenarios. One option is a single node installation on a very powerful machine with large capacity. A second is multiple node installation where there is one Archivematica pipeline, running on many (potentially virtual) machines. Another option is multiple installations, which run independently (perhaps one per department or workflow) but share archival storage. Finally, Archivematica allows for multiple independent installations, each with separate archival storage.

6. SUSTAINABILITY

The first production release signals to the Archivematica community of users that they can download and use the system as-is to complete their digital curation workflow. Subsequent releases will allow for an enhanced system that can continue to get better over time. More scalability testing will help to optimize production and, if necessary, select tools that perform better to accomplish micro-service tasks. The quality of Archivematica 1.0 was especially important since the open-source development model relies heavily upon community adoption and support.

There are several open-source development funding models, but two of the most pronounced are funding by a foundation or trust and crowdsourced funding led by a third party company or organization. In its beginning, it appeared as though Archivematica would be of the first type, funding largely by UNESCO. However, as the system evolved and Artefactual partnered with the City of Vancouver, it was clear that the project was destined to follow the latter model. Clients partner with Artefactual to fund the development of features that are in turn shared with the community. Distinct benefits of this model are its agility and variety of users and clients. Archivematica deploys agile development by setting release deadlines with a prioritized list of requirements [18], which puts pressure on Artefactual to release as much as we can during each release cycle so the community can evaluate changes and comment on their value. Artefactual makes it easy for the community to contribute via its user forum [19] and public issue management system [20].

The open-source development model encourages users to stretch their investments by pooling their technology budgets. This means the digital preservation community pays only once to have features developed, either by in-house technical staff or by third-party contractors like Artefactual. Archivematica project staff provide free community support and free software release management. All the software and documentation gets released under open-source (AGPL3) license and is offered at no cost, in perpetuity, to the rest of the user community.

This stands in contrast to a development model driven by a commercial vendor, where institutions share their own expertise to painstakingly co-develop digital preservation technology but then cannot share that technology with their colleagues or professional communities because of expensive and restrictive

software licenses imposed by the vendor. Commercial vendors benefit from the knowledge, time and money invested in open-source tools without contributing in-kind, or worse, selling the tools back to digital preservation colleagues in one form or another. The open-source model employed by Artefactual also stands in contrast to the "freemium" style open source business model, in which code is released while documentation or some other deployment necessity is withheld from all but paying partners.

As the community grows and contributes to the system, Artefactual can focus on building up a preferred provider network of trusted service providers. Preferred partners will be those contractors that can demonstrate their ability to provide users with high-quality support, and share Artefactual's open source values--that is, they will provide code completely free (AGPL3 license [21]) as a service to the archives and library community. A widening scope of service providers beyond just Artefactual will allow the Archivematica project to focus on innovation and moving forward for the benefit of all its users.

7. CONCLUSION

Thanks to a growing community of dedicated beta testers and client pilots, the first production version of Archivematica is a full-production digital preservation system, ready to be implemented, integrated with other systems, and developed further by Artefactual and community members. Baseline requirements for system scalability will continue to be stress tested over time to allow for enhancement and improvements as new features are added. User customization options allow for flexibility in repository workflows. Each new digital repository interface will be based upon a generic version and/or API, usable beyond the sponsoring repository. Format policy implementation, while being essential in staying current with preservation planning best practices, can have a broad effect in the larger digital preservation community, allowing for quantifiable data about normalization processes, successes and failures over time. Finally, a business plan that stays true to the ideals of the free software community will allow for new feature development and enhancements over time and nurture system sustainability.

8. REFERENCES

[1] Van Garderen, P. 2010. Archivematica: Using micro-services and open source software to deliver a comprehensive digital curation solution. *iPres Proceedings.* (Sept. 2010), http://www.ifs.tuwien.ac.at/dp/ipres2010/papers/vanGarderen28.pdf

[2] Mumma,C. and Van Garderen, P. 2012. The Community-driven evolution of the Archivematica project. *iPres Proceedings.* (Oct.. 2012), 164-171, https://ipres.ischool.utoronto.ca/sites/ipres.ischool.utoronto.ca/files/iPres%202012%20Conference%20Proceedings%20Final.pdf

[3] Bradley, K., Lei, J., Blackall, C. Towards An.Open Source Archival Repository and Preservation System (2007), http://www.unesco.org/webworld/en/mow-open-source/

[4] ISO 14721:2003. Space data and information transfer systems -- Open archival information system -- Reference model.

[5] Archivematica public wiki, OAIS use cases. (2009), https://www.archivematica.org/wiki/OAIS_Use_Cases.

[6] Van Garderen, P. 2010. Archivematica: Using micro-services and open source software to deliver a comprehensive digital curation solution. *iPres Proceedings.* (Sept. 2010), http://www.ifs.tuwien.ac.at/dp/ipres2010/papers/vanGarderen28.pdf

[7] Van Garderen, P. 2010. Archivematica: Lowering the Barrier to Best Practice Digital Preservation. *IS&T Archiving proceedings* (May 2010), http://www.imaging.org/IST/store/epub.cfm?abstrid=43770

[8] Mumma,C. and Van Garderen, P. 2012. The Community-driven evolution of the Archivematica project. iPres Proceedings. (Oct.. 2012), 164-171, https://ipres.ischool.utoronto.ca/sites/ipres.ischool.utoronto.ca/files/iPres%202012%20Conference%20Proceedings%20Final.pdf

[9] Archivematica public wiki, DSpace export and integration. (2012), https://www.archivematica.org/wiki/DSpace_integration, https://www.archivematica.org/wiki/DSpace_exports

[10] Open Archives Initiative. http://www.openarchives.org/

[11] Archivematica public wiki, CONTENTdm integration. (2012-2013), https://www.archivematica.org/wiki/CONTENTdm_integration

[12] Application Programming Interface (API). *Wikipedia*. (accessed April 20, 2013), http://en.wikipedia.org/wiki/Application_programming_interface

[13] Open Planets Foundation. http://www.openplanetsfoundation.org/

[14] CURATEcamp public wiki. http://curatecamp.org/

[15] Archivematica public wiki. Format policy registry requirements. (2012-2013), https://www.archivematica.org/wiki/Format_policy_registry_requirements

[16] Archivematica. Video of multiple processors. *Youtube.* (2012) https://www.youtube.com/watch?feature=player_embedded&v=lOZ-Kcw4DQs.

[17] Archivematica public wiki.Scalability testing documentation. (2011-2013), https://www.archivematica.org/wiki/Scalability_testing

[18] Archivematica public wiki. Development roadmap. https://www.archivematica.org/wiki/Development_roadmap:_Archivematica_1.0

[19] Archivematica user forum. https"//groups.google.com/forum/?fromgroups#!forum/archivematica

[20] Archivematica public issues list. https://projects.artefactual.com/issues/

[21] AGPL3 license. http://en.wikipedia.org/wiki/Affero_General_Public_License

Measuring Perceptions of Trustworthiness: A Research Project

Devan Ray Donaldson
University of Michigan
School of Information
3349C North Quad, 105 S. State St.
Ann Arbor, MI, 48109
devand@umich.edu

ABSTRACT
The digital curation and preservation community has long acknowledged that trustworthiness is a critical component of successful digital repositories. However, there is no known method to determine if or under what circumstances an end-user perceives a repository as trustworthy. While the research literature describes definitions, criteria, and certification processes that allow repository managers to assert trustworthiness under certain conditions, it does not adequately define, measure, or specify trustworthiness from the perspective of the end-user. This paper highlights traditional notions of trustworthiness in the context of the literature on digital repositories and explores trustworthiness from the end-user's perspective. The paper also presents an ongoing research project to: (1) investigate designated communities' perspectives on trustworthiness using focus groups, and (2) explore building, testing, and assessing an index to measure trustworthiness.

Categories and Subject Descriptors
H.1.2 [**Information Systems**]: User/Machine Systems – human factors, human information processing.

General Terms
Human Factors, Measurement, Reliability, Verification.

Keywords
Digital Curation, Digital Preservation, End-Users, Perceptions, Psychometrics, Trusted Digital Repositories, Trustworthiness.

1. INTRODUCTION
In the field of digital curation and preservation, repositories are asserted as "trusted" or "trustworthy" if they meet certain conditions encoded in best practices and standards. For example, the *Trustworthy Repositories Audit and Certification: Criteria and Checklist* (TRAC) is designed to measure and document the procedures and processes used to build and manage a repository and protect its content from corruption. The type of trustworthiness repositories that abide by these standards and best practices assert is very specific, having to do with a repository's ability to sustain itself and preserve digital materials [12].

Recently, empirical research studies have advanced our understanding of the extent to which end-users accept repositories as special domains of managed information. This type of work has also begun exploring the end-user experience of accepting the trustworthiness repositories assert by examining the factors that affect users' perceptions of trustworthiness [1, 2, 6, 15, 17]. These studies' findings disagree on whether the type of trustworthiness repositories assert is the same as the type of trustworthiness end-users of these repositories accept. All of this makes end-user trustworthiness perception a vibrant and interesting area of research.

This research project focuses more specifically on the notion of repository trustworthiness by taking the position that end-user perception of trustworthiness for individual documents or clusters of documents in a repository may affect perception at the repository level. The study also takes into account the idea that users can perceive trustworthiness in two ways: (1) as binary (e.g., trustworthy or untrustworthy), or (2) as a spectrum with a range of perceptions (e.g., more trustworthy or less trustworthy), based upon a wide range of factors. In addition, the study contextualizes end-user perception for a specific user group who uses specific types of documents from one repository, namely genealogists using marriage, death, birth, and family history records from the Washington State Digital Archives (WADA).

The purpose of this study is two-fold: (1) to investigate designated communities' perspectives on trustworthiness, and (2) to consider the extent to which trustworthiness is measurable as a construct of end-user perception for repository documents. This paper explicates details about focus groups and the method of index construction in an attempt to address both thrusts of the study's purpose. The focus groups will engage members of a designated community in conversation about their perceptions of trustworthiness for documents they have encountered while using WADA, and especially, how they develop those perceptions. Then, a multi-stage index construction process builds on those results and existing measures of trustworthiness to explore the extent to which a validated, multi-item index for assessing end-user perception of trustworthiness for repository documents can be obtained.

2. LITERATURE REVIEW

2.1 Asserting Trustworthiness
Examining the titles of significant international project reports and validation/certification programs demonstrates that, by and large, the digital curation and preservation communities conceptualize trustworthiness as a property that repository managers can assert over their repositories if they meet certain criteria. The *Trustworthy Repositories Audit and Certification: Criteria and Checklist* [12] specifies that if a repository adheres to specific criteria regarding organizational infrastructure, digital object management, and technical infrastructure, including security issues, it can be considered trustworthy. Consonantly, the *NESTOR Catalogue of Criteria for Trusted Digital Repositories* [10] delineates criteria for a repository's organizational framework, object management, infrastructure, and security that,

if met, result in repositories attaining trusted status. The *Digital Repository Audit Method Based on Risk Assessment* (DRAMBORA) [5] notes that if a repository identifies and properly manages threats to preserving digital materials, it can be considered trustworthy. The Data Archiving and Networked Services (DANS) – Data Seal of Approval (DSA) [7] outlines guidelines for the data producer, data consumer, and data repository that, if met, allow a repository to emboss an actual seal of approval on its website as an attestation of its trustworthiness. Similar to the other guidelines/standards, "[t]he seal of approval does not express any views regarding the quality of the data to be archived, but does regarding the provisions an archive has made to guarantee the safety and future usability of the data" [7, n.p.]. Often, the type of trustworthiness that repositories (which abide by these standards and best practices) assert focuses on a repository's ability to sustain itself and preserve digital materials.

2.2 Factors Affecting End-User Trustworthiness Perception

Despite differences in the type of repository, the type of content used, the type of user, and the purpose for using content from a repository, existing empirical research on trustworthiness perception for end-users of digital repositories suggests that a variety of factors affect their perception at both repository and document levels. Specifically, at the repository level, prior experience, institutional/organizational reputation, and third party endorsement are among the factors that affect end-user trustworthiness perception. At the document level, the author/creator/producer of the information, peer review, institutional/organizational reputation, a document or dataset's presence in a repository, and use purpose are among the factors that affect end-user perception of trustworthiness. In addition, findings vary regarding the extent to which repository level trustworthiness perception interacts with trustworthiness perception at the document level.

2.2.1 Factors Affecting End-User Trustworthiness Perception at the Repository Level

A user's prior experience with a repository is a factor that can help in determining the trustworthiness of a repository. St. Jean et al. [15] found that end-users base their perceptions of whether a repository is trustworthy on their prior experience with that repository. Specifically, the findings suggest that the more positive experiences end-users have with repositories, the more trustworthy they perceive those repositories to be. Likewise, participants in the Conway [2, p. 455] study perceived the repository they dealt with as trustworthy because of having "consistently positive experience in obtaining relevant, useful, and technically appropriate" content.

In addition to their own experience with repositories, users consider others' experiences as well, via the repository's reputation or track record. The CASPAR Consortium [1] found that users of curated digital objects rated the track record of a repository's ability to curate objects the most important factor among sixteen others in determining if a repository is trustworthy. Similarly, Yakel et al. [17] found that both archaeologists and quantitative social scientists mentioned institutional reputation as an important trustworthiness factor. Specifically, quantitative social scientists were twice as likely to mention the importance of institutional reputation as compared to archaeologists. Furthermore, novice quantitative social scientists were twice as likely to mention the importance of institutional reputation as expert quantitative social scientists.

Users' first-hand experiences are seemingly more important to their perceptions of trustworthiness than external factors like certification. While in their study, Ross and McHugh [13] took as axiomatic that certification is one marker that helps users determine the trustworthiness of a repository, subsequent studies that collected data from actual users of digital repositories show that third party endorsement might not be key in determining the trustworthiness of a repository. For example, the CASPAR Consortium [1] found that users of curated digital objects rated the fact that a repository has been validated by a toolkit such as DRAMBORA or TRAC and the fact that a repository has been validated by a domain-specific authority such as the Museums Documentation Association (MDA) among the least important factors in determining the trustworthiness of a repository. Similarly, in the Yakel et al. [17] study, only one quantitative social scientist cited seals of approval, a form of third party endorsement, as a factor that positively influences trustworthiness perception.

2.2.2 Factors Affecting End-User Trustworthiness Perception at the Document Level

Across multiple studies, the importance of the author/creator/producer of the content is an important factor for some end-users in determining the trustworthiness of repository content. For example, faculty, library staff, museum staff, undergraduate and graduate students in the St. Jean et al. [15] study were concerned about who created the content and why. People engaged in environmental planning including professionals employed by state, local, and federal agencies, representatives of environmental organizations and industry, concerned residents, and landowners in the Van House et al. [16, p. 340] study wanted not only to know who created the content, but to understand the extent to which the creator "followed the appropriate scientific practices" as part of their determination of the "trustability" of a dataset. Study participants also indicated that they needed to know the reputation of the content creator in order to determine the trustability of a measurement dataset. In contrast, in the Fear and Donaldson [6] study, awareness of a content creator's reputation was insufficient grounds for perceiving a dataset as trustworthy. According to the faculty members, postdoctoral fellows/researchers, staff scientists, and consultants in the study, some scientists with good reputations make "crap data" available while some other relatively unknown scientists create very trustworthy data.

Prior research suggests that if some end-users assume or know content has been subject to peer review, they will perceive that content as more trustworthy than they would otherwise. For example, in the St. Jean et al. [15] study, some faculty, library staff, museum staff, undergraduate and graduate students perceived institutional repository content as more trustworthy because they were under the impression that the content was subject to some sort of peer review process. In the Fear and Donaldson [6] study, the faculty members, postdoctoral fellows/researchers, staff scientists, and consultants were aware of the fact that not all the datasets in the proteomics repository had

been subject to peer review. In response, study participants actively sought out datasets that were associated with published articles and perceived those datasets as more trustworthy than other datasets that were unassociated with publications. The reason the study participants had such positive trustworthiness perceptions for datasets that were associated with publications was because, in the field of proteomics, both publications and their associated data are peer reviewed. In contrast to some of the other studies, the repository under investigation in the Van House et al. [16] study housed unpublished material and the study participants understood that the content had not been peer reviewed. Thus, participants did not rely on peer review to serve as a heuristic for perceiving content encountered within the repository as either trustworthy or untrustworthy, as some respondents in the St. Jean et al. [15] study assumed they could, or as some participants in the Fear and Donaldson [6] study actually could.

In the St. Jean et al. [15] study, faculty, library staff, museum staff, undergraduate and graduate students indicated that a repository's tie with an institution positively influences their perceptions about the trustworthiness of the content they find. They assumed that an institution would not allow information that was untrustworthy to be made available via the repository, because they assumed the institution would not jeopardize its own reputation by providing untrustworthy information.

For some end-users, the presence of a dataset in a repository serves as an indication of its trustworthiness. Fear and Donaldson [6] found that some proteomics researchers believe that a scientist's willingness to make his or her data available in a repository demonstrates that it is trustworthy enough to be used by others. These study participants subscribed to the idea that data producers would not willingly make untrustworthy data available because doing so could jeopardize a data producer's reputation.

Levy [9] first pointed out that the use to which digital documents will be put is an important consideration that should guide choices about digital preservation. Subsequent empirical research suggests that use purpose is moderated by end-user trustworthiness perception for documents preserved in a repository. For example, in the Fear and Donaldson [6] study, participants perceived some of the preserved datasets as trustworthy enough to replicate the analysis of the data creator, but those same data were not perceived as trustworthy enough to actually understand the biology behind the data, and were thus insufficient for that use purpose.

2.3 The Interaction of Repository and Document Level Trustworthiness Perception

Results vary regarding the extent to which repository and document trustworthiness perceptions interact. Conway [2, p. 455] found that, for his study participants, trustworthiness "ascends to the organizational level and, as a consequence, pervades the resources delivered digitally." In contrast, Yakel et al. [17, p. 11] found that "[t]rust in the repository is a separate and distinct factor from trust in the data." Taken together, the findings motivate a need for more research to better understand when repository trustworthiness perceptions affect document trustworthiness perceptions, and when they do not.

2.4 End-User Conceptualization of Trustworthiness

In the St. Jean et al. [15] study, repository end-users articulate their conceptualization of trustworthiness in a way that suggests it is multi-faceted for them. They interpreted "trustworthy" as comprehensive, factual, legitimate, professional, reliable, reputable, updated, and verifiable.

2.5 Summary of Literature as Motivation for Study

Taken together, the literature demonstrates that trustworthiness is central to justification for digital repositories, but it has only been asserted as a concept. Trustworthiness has not been defined in a way that is amenable to verifying its presence or absence in a repository context from the end-user's perspective. The research on end-users has identified factors that affect their perception of trustworthiness at both repository and document levels. These findings provide insight into assumptions end-users make about the type of trustworthiness repositories assert. Existing empirical research also suggests that end-user conceptualization of trustworthiness is multi-faceted [15]. For any repository, understanding how their designated communities conceptualize trustworthiness is necessary, as is measuring trustworthiness perception based upon that conceptualization. This paper describes the development of a composite measure for assessing designated communities' concept of trustworthiness.

3. A RESEARCH PROJECT

3.1 The Washington State Digital Archives (WADA) as the Primary Site of Study

The research study centers on end-user perception of trustworthiness for preserved documents found in digital repositories. In order to conduct the investigation, the Washington State Digital Archives (WADA) will serve as the primary site of study for five reasons. First, WADA is a highly utilized digital cultural heritage resource. Approximately 500,000 people visit WADA per year with thousands of unique visitors per month. Second, WADA has a strong and explicit mission statement, which focuses on making preserved digital information accessible to users. Third, WADA has had a great deal of success in administering web surveys to their users. Fourth, the author can access WADA data relatively seamlessly because of an established relationship with WADA administrators. Fifth, in action and deed, WADA is a Trusted Digital Repository (TDR) that abides by leading practices and standards for organizational infrastructure, digital object management, and technical infrastructure, including security issues, consistent with TRAC specifications, despite not being formally certified as a TDR as of April 2013.

This study focuses on genealogists, who represent WADA's largest designated community (personal communication with WADA staff, March 8, 2013). Also, based on WADA's download statistics, genealogical records are among WADA's most highly downloaded records. For these designated community members, most of the records they utilize are digitized records available for download in JPEG format accompanied by transcriptions. In some cases, only the digitized record is available, and in other cases, only the transcribed version is available.

3.2 Focus Groups

Before attempting to build, test, and assess an index to measure the construct of trustworthiness, one must understand designated communities' perspectives on trustworthiness. According to Stewart and Shamdasani [14], one of the uses of focus groups is to learn about how respondents talk about a phenomenon. The research study will use focus groups to collect data from genealogists to understand their perspectives on trustworthiness.

To recruit participants, WADA staff will forward a description/invitation for the study to users who they know have a track record of using WADA. Those interested will utilize the contact information provided in the study description to call or email the author directly and finalize arrangements for participating in the focus groups. The target size of each focus group is six to eight participants.

Each participant will take a paper-based pre-survey before the focus group begins. It will include the following questions/prompts:

1. On average, how frequently do you use the Internet?
2. How strongly do you agree with the following statement: In general, I trust information I find on the Internet.
3. In the last year, how frequently have you used the Washington State Digital Archives?
4. What is your primary reason for visiting the Washington State Digital Archives?
5. How strongly do you agree with the following statement: I usually find the documents I'm looking for when using WADA.
6. How strongly do you trust the documents you find when using the Washington State Digital Archives?
7. How satisfied are you with the way the Washington State Digital Archives displays documents?

Question 1 engages participants' Internet usage. Question 2 examines participants' disposition to trust information found on the Internet broadly speaking. Question 3 is useful for understanding the extent to which the study participants have a track record of using WADA. Question 4 focuses on participants' primary reason for using WADA. Questions 5-7 investigate participants' experiences with and perceptions of WADA documents. In addition, the pre-survey includes two demographic questions related to participants' age and gender.

To maximize breadth and depth of discussion in the focus groups, the author will ask the following open-ended questions/prompts:

1. Discuss the nature of the documents you use when using WADA and your purpose(s) for using them.
2. Discuss your perceptions of trustworthiness for the documents you find using WADA.
3. How would you describe a document you found in WADA that you think is trustworthy?
4. Under what circumstances would you question the trustworthiness of a document you encountered while using WADA?
5. Card-sorting exercise.

Question 1 is designed to be an "icebreaker" question, which, according to Stewart and Shamdasani [14], is how any focus group should begin. The question engages participants' use of WADA, including their purposes. Questions 2-4 specifically engage trustworthiness in the context of WADA and for documents encountered within it. Question 5 is a card-sorting exercise in which participants will break into pairs and sort potential trustworthiness perception attributes into three piles in terms of how important they think they are for the documents they use: important, somewhat important, and not important. After participants complete the card-sorting exercise, we will discuss how and why each pair grouped the attributes the way they did.

The focus groups will take place on-site at WADA and be videotaped. Each focus group will last for approximately an hour and a half. The resulting data will be transcribed and analyzed using nVivo 9.0. Overall, the focus groups will inform our understanding of these designated community members' perspectives on trustworthiness, including their conceptualization of the construct, laying the groundwork for the next phase of the research project.

3.3 The Index Construction Process

There are four steps to an index construction project, including: (1) construct definition, (2) generating an item pool, (3) designing the index, and (4) full administration and item analysis [4].

3.3.1 Implementing Step 1 – Construct Definition

Step 1 involves completion of three tasks related to defining trustworthiness. First, development of a brief definition of trustworthiness, including its scope and any subcomponents that are to be included. Second, further development of the definition of trustworthiness by drawing upon existing definitions from relevant research studies and theoretical literature published in digital preservation and curation, communication studies, information science, and web credibility domains. Third, operationalization of the construct definition of trustworthiness by: (1) considering the different types of questions or rating scales to which study participants can respond, and (2) asking oneself what kinds of responses would be clear indicators of the respondents' levels or amounts of perceived trustworthiness.

3.3.2 Implementing Step 2 – Generating an Item Pool

To implement Step 2, a number of tasks will be completed related to generating an item pool for the construct of trustworthiness. Any existing instruments that measure trustworthiness will be examined. Items from those instruments will be selected as a starting point for the initial item pool. If these instruments do not exist, related instruments will be examined, which may contain items that are acceptable for inclusion. If no items in existing or related instruments are appropriate, the researcher will create them. In addition, ideas for items will be gathered from reviewing the literature on trustworthiness. Items will also be generated from members of WADA's largest designated community (i.e., genealogists) who will be asked to articulate, during the focus groups, adjectives to describe documents they think are trustworthy. These trustworthiness attributes will be reviewed to assess the extent to which they compare or contrast with: (1) items found in the literature, and (2) items experts recommend.

By manual inspection, pretesting with a small sample of respondents, and conferring with experts, a host of issues that must be considered during Step 2 will be addressed, which include [4]:

- ensuring each item expresses only one idea
- avoiding lack of colloquialisms, expressions, and jargon
- ensuring the reading difficulty matches the reading level of respondents
- ensuring the items match the specificity of trustworthiness
- ensuring that what the items have in common is trustworthiness and not merely a category
- ensuring that the item pool consists of an exhaustive list of items that appear to fit trustworthiness
- avoiding exceptionally lengthy items
- making items as short and as uncomplicated as possible.

Expert involvement will play a major role in Step 2. The researcher will assemble a panel of trustworthiness experts to evaluate the entire initial item pool. In a self-administered web survey, the experts will be provided the construct definition developed during Step 1 and then they will be provided with the initial item pool. The survey instructions will ask the experts to rate each item with respect to trustworthiness according to the following designations: essential, useful but not essential, or not necessary. Experts' responses will be analyzed by computing a Content Validity Ratio (CVR) for each item [8]. For purposes of this study, all items that have positive CVRs will be retained for the trustworthiness item pool. In addition, the instrument will ask experts to: comment on individual items as they see fit, evaluate the items' clarity and conciseness, point out awkward or confusing items, suggest alternative wordings, and suggest additional items.

3.3.3 Implementing Step 3 – Designing the Index

Step 3 involves a number of activities related to the format of the index, including: selection of response categories and choices, writing item stems, and writing instructions. This step also involves pretesting.

The definition of trustworthiness developed during Step 1, coupled with the researcher's understanding of the literature on index construction for Step 3, will guide selection of response categories and choices.

Following the recommendation of authors on the topic of index construction [4], the researcher anticipates choosing seven response choices. This odd number of choices will allow respondents the option of neutrality if particular items are neither important nor unimportant to them. In addition, seven response options will allow a greater level of granularity with respect to the resulting data. The various gradations of importance will enable the researcher to discover if and to what degree items are important or unimportant to end-users.

Item stems will be written with the construct of trustworthiness in mind. As well, item stems will be written with the response categories in mind; they will be made as clear, concise, unambiguous, and concrete as possible.

The item pool instrument will be administered as a web survey because WADA end-users are geographically dispersed. Thus, administering the item pool instrument as a web survey would make it much more feasible for respondents to participate in the project.

Step 3 will also include informal pretesting and formal pilot testing of the draft instrument including cognitive interviews. For the informal pretesting, members of the Archives Research Group (ARG) at the University of Michigan School of Information will be recruited and emails will be sent out on student listservs to recruit Master's and Ph.D. students. Each participant will be asked to indicate if any items are ambiguous or confusing, or if they feel any items cannot be rated along the response categories and choices provided by the instrument. The index will be revised on the basis of participants' feedback. For the formal pilot testing, the researcher will travel to Washington to administer the index to a small group of actual WADA end-users, which WADA staff will help identify and recruit. Each respondent will complete the instrument in a private setting in the WADA Reading Room while the researcher is present. Each respondent will be asked to think aloud. Similar to the pretest participants, the pilot test participants will also be asked to indicate which items are ambiguous or confusing, and which items cannot be rated along the instrument's response categories and choices. This type of evaluation will be used to identify items that are not clear, items that are being interpreted in ways that are different from how they were intended, as well as instructions that are vague or ambiguous.

3.3.4 Implementing Step 4 – Full Administration and Item Analysis

To implement Step 4, the item pool generated during Step 2 and pretested in Step 3 will be administered as an instrument and item analysis and factor analysis will be conducted. After each statistical test, the results will be used to further improve the instrument, deleting or revising any items that are not contributing to its quality. This iterative process will continue until the instrument is of sufficient quality.

The sample population for this study will be actual WADA end-users from its largest designated community. The instrument will be administered to these users as an intercept survey [3]. For example, every 200th visitor will receive a pop-up invitation to participate in the study. This form of probabilistic sampling (i.e., systematic sampling) is a practically viable way of making sure actual WADA end-users participate randomly in the survey. In addition, screening questions will enable the researcher to identify those participants who self-report as genealogists.

The number of participants for the study will be a function of the number of items in the instrument. Specifically, the researcher will follow Nunnally's [11] recommendation of between 4 and 10 participants per item.

After administering the instrument to a sample of WADA end-users, several characteristics of individual items will be evaluated.

During item analysis, the researcher will examine item-scale correlations, item variances, and item means.

To assess intercorrelation, the researcher will compute item-index correlations for each item. Corrected item-index correlation will be computed, rather than uncorrected item-index correlation because the latter could inflate reliability [4]. The researcher will also assess item variances by examining the range of responses for each item, anticipating retaining response items broadly, per DeVellis's [4] recommendation.

Both item-index correlations and coefficient alpha will be used to choose items for an index. Depending on the findings, a series of steps may be taken, such as deleting some items, checking alpha, deleting more items, and rechecking alpha, until a final set of items is chosen.

During factor analysis, varimax rotation will be conducted and scree plots will be generated, paying close attention to those factors which have the highest eigenvalues. The results of the factor analysis will be examined to see if they make logical sense or make sense in light of existing theory.

Although item-index correlations may be used from a statistical perspective to understand the extent to which certain items relate and could therefore be useful for measuring trustworthiness, results of statistical analyses will not be relied upon solely to build the index. The researcher will consider theoretical and practical understanding of the items in light of statistical calculations prior to finalizing conclusions about what is being measured.

After administering the item pool as an instrument and completing item analysis and factor analysis on the data, a modified version of the index may be administered to another sample of WADA end-users, performing item analysis and factor analysis on the resulting data. The goal of Step 4 is to achieve an internally consistent and logically sensible instrument. Administering a modified version of the index may or may not be necessary. It will be dependent on results of the first item analysis and factor analysis; the researcher's subjective judgment; and consultation with specialists with expertise in researching trustworthiness regarding whether the items the statistics suggest correlate make sense to be considered together.

4. CONCLUSION & SIGNIFICANCE

The research project is significant because it attempts to answer one of the most important questions in the digital curation and preservation research domain, "When is a repository trustworthy?" Specifically, the study is designed to:

- explore what trustworthiness means to actual end-users
- operationalize trustworthiness, and
- measure trustworthiness.

There is value in conducting this study regardless of the outcome. If completion of Steps 1 through 4 result in an internally consistent and logically sensible instrument, then the instrument's mere existence validates claims by researchers that complex constructs cannot be measured reliably using one item [8], and that trustworthiness is no exception. Further, the instrument will provide a specific composite operationalization of the construct of trustworthiness which could be tested for validity in various contexts, such as with documents found in TDRs besides WADA. If completion of Steps 1 through 4 does not result in an internally consistent and logically sensible instrument, substantial insight will be discovered concerning the challenges to measuring trustworthiness. This specific outcome would suggest that more conceptual work needs to be done on trustworthiness.

Ultimately, investigating and measuring trustworthiness is precisely the type of work digital curation and preservation researchers need to conduct to understand and monitor designated communities' perceptions of trustworthiness for repositories. This paper describes an ongoing research project with a methodology to administer such a process. The timeframe for this project is approximately one year from start to finish and is broken into four main phases: Step 1 (February 1, 2013 – April 30, 2013), Step 2 including focus groups (May 1, 2013 – July 31, 2013), Step 3 (August 1, 2013 – August 31, 2013), Step 4 (September 1, 2013 – November 30, 2013), data analysis and report writing.

5. ACKNOWLEDGMENTS

This material is based upon work supported by a Graduate Student Research Grant from the Horace H. Rackham School of Graduate Studies at the University of Michigan. The author would like to thank Paul Conway, Kathleen Fear, William Jacoby, James Lepkowski, Soo Young Rieh, Charles Senteio, and Elizabeth Yakel for providing comments on previous drafts of this paper, as well as Terence S. Badger and the staff at the Washington State Digital Archives for their help and support.

6. REFERENCES

[1] CASPAR Consortium. 2009. Report on Trusted Digital Repositories. Technical Report.

[2] Conway, P. 2010. Modes of seeing: Digitized photographic archives and the experienced user. *American Archivist* 73(2): 425-462.

[3] Couper, M. P. 2000. Review: Web surveys: A review of issues and approaches. *Public Opin. Q.* 64(4): 464-494.

[4] DeVellis, R. F. 2012. Scale development: Theory and applications. Thousand Oaks, Calif.: SAGE.

[5] Digital Curation Centre and Digital Preservation Europe. 2007. *DCC and DPE Digital Repository Audit Method Based on Risk Assessment*, v.1.0. Technical Report.

[6] Fear, K. and Donaldson, D. R. 2012. Provenance and Credibility in Scientific Data Repositories. *Archival Science* 13(1): 55-83.

[7] Harmsen, H. 2008. Data seal of approval - assessment and review of the quality of operations for research data repositories. *Proceedings of the Fifth International Conference on Preservation of Digital Objects* (St. Pancras, London, Sept. 29-30, 2008). iPres'08. 1-3.

[8] Kim, Y. 2009. Validation of psychometric research instruments: The case of information science. *J. Am. Soc. Inf. Sci. Technol.* 60(6): 1178-1191.

[9] Levy, D. M. 1998. Heroic measures: Reflections on the possibility and purpose of digital preservation. *Proceedings of the Third ACM Conference on Digital Libraries* (Pittsburgh, Pennsylvania, USA). DL'98.

[10] NESTOR Working Group on Trusted Repositories Certification. 2009. Catalogue of criteria for trusted digital repositories Version 2. Deutsche Nationalbibliothek: NESTOR Working Group.

[11] Nunnally, J. C. 1978. *Psychometric theory*. New York: McGraw-Hill.

[12] RLG-NARA Digital Repository Certification Task Force. 2007. Trustworthy repositories audit and certification: Criteria and checklist. OCLC and CRL.

[13] Ross, S., & McHugh, A. 2006. The role of evidence in establishing trust in repositories. *D-Lib Magazine* 12(7/8).

[14] Stewart, D. W., and Shamdasani, P. N. 1990. Focus groups: Theory and Practice. Applied Social Research Methods Series; v. 20. Newbury Park, Calif.: Sage.

[15] St. Jean, B., Rieh, S. Y., Yakel, E., and Markey, K. 2011. Unheard voices: Institutional repository end-users. *College & Research Libraries* 72(1), 21-42.

[16] Van House, N., Butler, M. H., and Schiff, L. R. 1998. Cooperative Knowledge Work and Practices of Trust: Sharing Environmental Planning Data Sets. *Proceedings of the 1998 ACM Conference on Computer Supported Cooperative Work* (Seattle, Washington, USA). CSCW'98.

[17] Yakel, E., Faniel, I., Kriesberg, A., and Yoon, A. Forthcoming. Trust in Digital Repositories. *International Journal of Digital Curation*.

A Framework for Automated Verification in Software Escrow

Elisabeth Weigl
SBA Research
Vienna, Austria
eweigl@sba-research.org

Johannes Binder
SBA Research
Vienna, Austria
jbinder@sba-research.org

Stephan Strodl
SBA Research
Vienna, Austria
sstrodl@sba-research.org

Barbara Kolany
ITM Münster
Münster, Germany
barbara.kolany@uni-muenster.de

Daniel Draws
SQS Research
Cologne, Germany
daniel.draws@sqs.com

Andreas Rauber
Vienna University of Technology
Vienna, Austria
rauber@ifs.tuwien.ac.at

ABSTRACT

If a business is in need of customized software it often orders it from a third party developer. This can lead to a dependency on this developer regarding maintainability and development of the product. Software Escrow offers a mitigation to this as it includes a trustable escrow agent in the business relationship. The agent is responsible for depositing all material that is needed to develop the software, like source code, documentation, and licenses for software artifacts. If a predefined trigger event occurs, the agent is obliged to hand out the objects to the customer. Thus the material needs to be of a quality that allows the customer to further maintain and develop the software. To guarantee this, all artifacts deposited are verified for their maintainability. As this verification is a time consuming and costly factor, we propose a Technical Software Escrow Framework that supports the reviewing process by highlighting parts of the software that can pose a problem regarding their maintainability. We also analyze an exemplary use-case software to show the applicability of our framework.

Keywords

Software Escrow, Software Quality, Evaluation, Verification Framework, Case Study

1. INTRODUCTION

For their daily work, businesses are in need of customized software. Thus they order the development and customization of software from other businesses, which commonly sell them a license for its usage. For the customer this then represents an asset of value, as he uses it for his day-to-day business. In order to adopt and adjust the software or to add new features, usually a service and maintenance contract is set up with the developer. This introduces a high dependency on the developer. In case he goes bankrupt or refuses to maintain the program, the customer will be negatively affected or in the worst case sustain severe financial effects. As common software licensing only includes the object code and not the sources of the software, the customer does not have access to the source code and thus will not be able to further develop or fix the software.

Software Escrow offers a mitigation to this scenario by placing a trustable party between the IT partner and his customer. The material relevant for the software development is deposited at the agent. To be able to further develop the software, it is important that the material gets checked. The agent is responsible for this verification and the subsequent storing for later re-use. During the depositing process, he has to ensure the physical security of the material. In case a trigger event occurs (e.g., bankruptcy), he is obliged to hand out the material to the customer, who wants to develop and maintain the software.

A successful escrow has several considerations to take into account. It has to be legally ensured that the future developing party has the rights for development, e.g., they have the right to use the source code and the libraries. These points are agreed on in the escrow contract, which is an extension to the commonly used license and maintenance contract and which has to be aligned with both of them. Other parts of the agreement involve decisions on the materials to be deposited, notification obligations, and trigger events that entail the release of the materials. With the trigger events clearly specified, the escrow contract helps to quickly release material and avoid delays in the procedure and legal uncertainties.

From a technical point of view, the escrow agent has to verify the completeness and evaluate the quality of the material relevant to the software project put into escrow, according to the agreements made in the contract. Completeness of material is needed because missing software artifacts can prevent a developer from maintaining the software [10]. Cur-

rent escrow approaches only focus on material consisting of source code and documentation, without detailed verification (e.g., only virus check of the data). Software projects, however, consist of more than that. Different artifacts like compilers, test scripts, external resources like Web services, or databases are also part of a software development project.

Maintainability of the deposited software is an important indicator for future maintenance and development processes and thus also has to be evaluated. Up until now maintainability of the deposited material is not promoted to be considered in escrow agreements. Standards like [6] only propose quality checks that do not evaluate maintainability comprehensively, including tests for e.g., readability of the data, random samples of the documentation, virus-free data, or compilation. These tests do not check all artifacts relevant for software projects.

The check for completeness and quality of the deposited material has to be done by a reviewer. A manual review conducted by the escrow agent requires sustainable effort and causes high costs. To increase the efficiency of the reviewer we developed a framework to support the verification process with automated artifact analysis. Thus we propose a Technical Software Escrow Framework implemented in Java that supports a manual review with automatic checks of the deposited software development project. For this purpose it pre-screens all artifacts, performs automatic checks and assessments, and highlights parts of the software project that need further examination by the reviewer (e.g., complex classes with minimal documentation).

The presented framework extends Software Escrow by Digital Preservation aspects of software development projects, like identifying dependencies to external resources of software such as Web services, which is needed to ensure long term availability of the service and its functionality. With this we introduce a framework capable of supporting the execution of Software Escrow by supporting manual evaluation actions of the reviewer with automatically executable processes and thus reducing the time needed for verification.

In this paper we will explain the different steps of Software Escrow and the technical verification in detail. With an exemplary use case, based on a Java open source project, we will go through the evaluation process and show the applicability of our framework. We will start with the related work on software quality important for Software Escrow in Section 2. An overview of Software Escrow and the escrow process follows in Section 3. An evaluation of our use case can be found in Section 4. In Section 5 we summarize the lessons learned and give a conclusion.

2. RELATED WORK

From a technical point of view the CEN Workshop Agreement 13620-5 - *ESCROWGUIDE* [4] offers a comprehensive information on Source Code Escrow and will be the basis for our investigation of Software Escrow. It comprises of five different parts: introducing Software Escrow, the view for each of the participants (developer, customer, agent), and one focusing on the audit process. Concerning technical aspects, the guide for developers [5] is the most interesting part for all parties regarding setting up a proper escrow contract. It describes what material to deposit, which will be necessary for our completeness check, where to put the escrow process in the software life-cycle, and gives an overview of the legal considerations for the developer.

As a theoretical concept for verification of the deposited material, the Escrowguide dedicated to the escrow agent [6] mentions three different levels. A *standard verification* only verifies the readability of the data, its completeness, or random samples of documentation. The *full verification* involves practical verification methods, including a compilation of the program and a test for functionality of the software. A few checks require the assistance of the software owner or client as well, which increases the effort for the affected parties. The third verification, the *bespoke verification*, may include tests from the standard or full verification together with additionally agreed tests. As not everything needed for the development of the software project is checked in the full verification, maintainability related checks can be agreed on here. In practice, a verification is done in more detail, including for instance the comparison between the compiled deposit materials and the executables running at the customer's site [13], or simulating a release event scenario [19].

A general introduction to the difficulties that arise can be found in [10]. It argues that the benefits of escrow do not compensate the time, legal fees, and other resources spent. We focus on the statement mentioned there that the material escrowed often is not usable after releasing it and try to approach this by extending common quality measurement methods.

Over time different standards were developed to describe and classify software quality. Whereas the ISO 9126 [15] set six main quality objectives, its successor and the current standard ISO 25010 [14] defines two quality models: one for quality in use, with five characteristics that relate to the outcome of interaction when a product is used in a particular context; the other for product quality, comprising of eight characteristics that relate to static properties of software and dynamic properties of the computer system. Important for Software Escrow are the two quality in use attributes portability and maintainability, which are in the focus of our framework.

For the quality tests we therefore focused on metrics that indicate maintainability and portability. Cyclomatic Complexity can be used to determine maintainability of source code. It was developed by Thomas J. McCabe in 1976 [17], based on the idea that humans can understand source code only until a certain amount of complexity of the code is reached. Instead of looking only at the syntactic elements in it, source code is seen as a directed graph with nodes and edges, nodes representing commands and edges representing direct connections between commands. According to McCabe a "reasonable, but not magical" [17] upper limit for the cyclomatic complexity is ten. Our framework will not stick to this number but we will compare our results to other popular open source projects. Related measures are Halstead's software metrics [9], including metrics like Program Volume, Difficulty, and Effort-To-Implement. Contrary to the cyclomatic complexity proposed by McCabe these met-

rics are based on lexical measures. Our framework uses this measurement to give an indication of understandability of the source code.

Regarding the measurement of quality in the source code comments, there are different metrics we used in the framework. The first one is comment density, which calculates the percentage of comments compared to lines of code and which we will combine with Cyclomatic Complexity to propose a new measurement . Arafat and Riehle [1] found that the average comment density in over 5000 successful open source projects was 18.67%.

As a second option to evaluate documentation quality, our framework also checks language for consistency and grammatical errors, which can make text hard to understand. Determining the language of a text and thus categorizing comments can be done with the usage of n-grams, as described by Cavnar and Trenkle in [3]. To proof text for correct spelling and grammar can also be an essential task in text analysis. Part-of-speech syntactic patterns as mentioned in Heyer et al. [11] or word sequence patterns that are compared to entries in error corpora as described in [18] can support this process.

Regarding the legal perspective of Software Escrow, [20] describes types of escrow agreements and release conditions. In [21] a short overview of the legal and technical aspects is presented. A detailed discussion of legal aspects regarding Software Escrow can be found in [12]. In this work we will highlight the most important aspects that have to be taken into account when setting up an escrow agreement.

3. SOFTWARE ESCROW

The subject of Software Escrow is a software produced by a developer for a customer and thus it refers to contractual agreements about the deposit of materials relevant for said software at a neutral third party. In case a contractually recorded trigger event occurs, the third party is obliged to hand over all materials to the customer.

Software Escrow agreements involve three parties:

- the **customer**, who has a need for a software in his business, wants to ensure that he is able to use the software for a longer time, and secure his investments in it
- the **software developer**, who makes the compiled object code available to the customer and hands over the sources and all other necessary artifacts to the escrow agent
- the **escrow agent**, who is responsible for depositing the material and releasing it, and who has to verify that the submitted material meets the requirements as contracted, e.g., that all objects are available, accessible, and fulfill specified quality measurements

Figure 1 shows an illustration of the relationship between the parties.

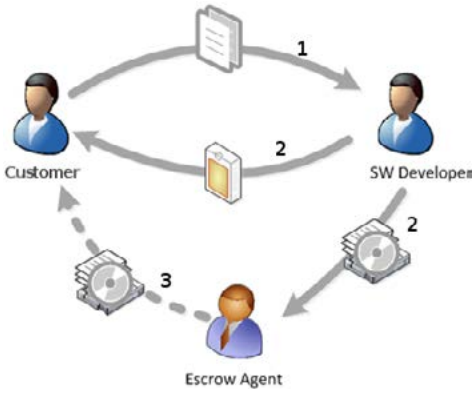


Figure 1: Relationship between escrow parties

There are technical and legal issues to be considered in a Software Escrow. From a legal point of view the contract needs to specify the obligations and rights of all three parties, the material to deposit at the agent, and the release events and procedure. It is important to exactly specify the events that entail the release of the deposited material to avoid legal uncertainties. The verification procedure and its success criteria also need to be stipulated. Licenses and rights to the material are as well part of the contract.

From a technical point of view, the completeness and quality of the deposited material have to be examined. The completeness of the material is crucial when the software has to be enhanced or maintained later. Thus all artifacts necessary for development have to be identified. This includes source code, libraries, compiler and compile instructions, test data, databases, and documentation amongst others. Availability of external dependencies of the software is necessary for preserving the functionality. Thus it is important to identify all external dependencies, like Web services or binaries used, in order to preserve them and therefore ensure the full functionality of the software over time.

Other technical considerations deal with the software's quality. Not mistakes or bugs are in the focus of Software Escrow, as they are part of the functional quality and thus part of an acceptance test by the customer. Maintenance aspects are important when depositing artifacts. All material has to be of a quality that ensures that it will be useable again. This also includes supplementary material needed for understanding certain artifacts, like documentation. These considerations are key drivers for our framework and will be explained in detail in Section 3.2, where the material is evaluated.

The escrow process can be divided into three phases: planning, execution, and redeployment. The focus of the Software Escrow planning phase (Section 3.1) lies on drafting the escrow agreement. The main task of the execution phase (Section 3.2) is the validation of the material against contractual requirements and its safe storage, as well as repeating those steps for each new version and update. The redeployment phase (Section 3.3) has to ensure the quick release of the deposited material once a contracted trigger event occurs.

3.1 Planning Phase

The planning phase is the first step in the escrow agreement and focuses on establishing an escrow contract. Preparations for an escrow contract should already be considered during the licensing contract negotiations. As certain costs are associated with setting up a Software Escrow, the first step should be an assessment of financial and business impacts if the software is unavailable. Then an appropriate escrow agent has to be selected. He has to be trustworthy for both parties. In the past a lawyer or notary was commonly chosen, however they often did not provide the necessary technical background needed. Nowadays specialized Software Escrow agents have been established, who are able to provide the knowledge needed for escrow as well as an appropriate technical infrastructure for evaluating and storing the deposited material.

In this phase the escrow agreement is composed and all parts of the contract are agreed on. The deposit material and its quality requirements have to be specified in the contract. As software can be a custom-made product, the artifacts needed for the deposit have to be specified for each project, including (based on [8]):

- Source code (including source code and libraries)
- Intellectual Property (especially licenses for different software components)
- Documentation (system and user documentation)
- Test environment (test cases, test scripts)
- Design environment (especially design models)
- Build environment (compilers, runtime environments, configuration files)
- Applications (databases or binary files that are used by the software)

Regarding the quality of these artifacts, the escrow agreement includes certain thresholds that have to be fulfilled. On the one hand this can be numerical boundaries, like compliance to a certain maximum source code complexity, and on the other hand the check for the fulfillment of requirements needed for immeasurable artifacts, like the requirements for a documentation of sufficient quality. The deposit procedure, including deadlines for the deposit, and the method of verification to fulfill the stipulated quality goals are agreed on as well [16].

The alignment of the licensing and maintenance contract with the escrow agreement (e.g., the specification of maintenance obligations) needs to be done in this phase as well. The escrow contract also needs to specify the transfer of rights, e.g., the allowance to use the source code or libraries needed for the maintenance of the software project. Rights exceeding the limitations of the original software contract, like commercial distribution of the program by the former customer, will have to be agreed on in the escrow contract separately.

3.2 Execution Phase

The execution phase is the second phase of the escrow process and depicted as the second step in Figure 1. It includes the deposit of the software, its verification, and the safe storage at the escrow agent, as well as the delivery of the program to the customer. As updates and new versions are released for the software, this process will be done repeatedly: With every update delivered to the customer, the material at the escrow agent has to be updated, verified, and deposited again. First the software, respectively a binary version of it, is delivered to the customer. At the same time the software development project and all its materials necessary for developing and maintaining the software are handed over to the escrow agent, where they get verified. If the verification is successful, the material gets stored safely, otherwise it gets rejected and the developer has to re-submit a revised version. These procedures have to be defined in the escrow contract.

The verification of the software has two main purposes: to ensure the completeness of the software development material and to verify the quality of each artifact. Both are necessary to guarantee the maintainability of the software once it gets handed out to the customer. The completeness check has to verify that each artifact agreed on and listed in the escrow agreement is part of the deposited materials. The quality evaluation includes verifications of the artifacts, like the quality of the documentation or that the sources do not exceed a predefined value for complexity. Each artifact has to be analyzed for its level of quality as specified in the escrow contract. To support this time-consuming process we developed a technical framework in Java that partly automates the verification process. It analyzes the artifacts and reports back to the reviewer those parts of the software that do not reach the required level of quality.

Our framework contains an extendable evaluation part with which various measurements can be conducted. It builds on the design of a Software Quality tool, which includes different static code analysis tools like Checkstyle[1], FindBugs[2] or PMD[3]. These support our maintainability evaluation because they are able to find source code sections that, e.g., contain code layout issues or flaws like unused variables that can make the source code difficult to understand. A tool combining these code analysis programs and different statistic code measurements is Sonar[4], an open source platform for continuous quality inspection, that forms the basis for our technical framework. Sonar supports the analysis of programs in several languages. With its client-server model the analysis can be run on a local system and the server provides different check modules for the client [2]. A project's quality is measured using metrics and rules, resulting in numerical values and violations, respectively. Sonar provides many measurements out of the box but can also be extended by integrating custom plugins. A description of the plugins developed for our Software Escrow scenario can be found below.

[1] http://checkstyle.sourceforge.net
[2] http://findbugs.sourceforge.net
[3] http://pmd.sourceforge.net
[4] http://www.sonarsource.org

In a Software Escrow scenario, the escrow agent first configures the framework according to the requirements agreed on in the escrow contract. The framework then processes the artifacts. Once it has finished it presents the reviewer with an overview of its findings, classifying the results according to their impact on the quality.

The following categories of quality checks related to maintainability, used to determine the quality of the deposit material, were implemented in our framework:

Completeness of artifacts. A reliable way to ensure completeness of the deposited source code material is to rebuild the software. Our prototype executes a build script and reports errors that may arise when doing so. The existence of other artifacts agreed on in the contract, such as additional documentation or specifications, can be assured either manually or using automatic checks as part of the build process.

Consistency of sources and released binary. The software put into escrow has to be the same as the one delivered to the customer. To verify this, the sources at the escrow agent have to be built and compared to the binaries delivered to the customer. Our implementation checks if the output generated by building the software matches a provided set of reference artifacts.

Quality of documentation. Documentation about the software development project is required to understand considerations and decisions made during the design and development phase in natural language. It is especially important if the software in question has to be maintained and possibly enhanced at an unknown time in the future because the programmer needs to understand the structure and design of the software. Thus the documentation has to be adequate, easily readable, and easily understandable. The following considerations apply to source code comments as well as additional documentation and specification, like architecture descriptions, requirement documents, manuals, etc.

One aspect that affects understandability is the language that has been used for the documentation. It needs to be ensured that all documentation is available in the agreed language. Our implementation detects the language of comments and reports if unexpected languages are found. To do so, the comments are extracted using SSLR[5] and analyzed using the Java Text Categorizing Library[6], which uses an algorithm based on n-grams [3]. To minimize the number of false positives, short comments can either be ignored or checked using a word list.

Also spelling, grammar, and other errors in documentation influence readability. Our implementation uses the text proof tool LanguageTool [18] to find issues of various categories like misspellings, wrong grammar, uncommon phrases, etc. in comments. It is possible to ignore specific issue types to filter frequently occurring mistakes that do not influence the readability, e.g., multiple whitespace characters, caused by specific formatting styles of comments. For each document the ratio of words compared to the number of issues detected in the comments is determined. This gives an overview of documents containing proportionally more errors than others.

Quality of source code. Software metrics can be used to assess the quality of the software, i.e., maintainability. An adequate level of quality is required for further development of the software. Sonar already implements a number of metrics for quality verification that can be used for our maintainability approach, such as the Cyclomatic Complexity. As mentioned in Section 2, Halstead's software metrics are a similar measurement method. As they are not implemented in Sonar, we provided a plugin for calculating the Halstead's software metrics Difficulty, Effort, Volume, Time to Program, and Bugs Delivered. For object oriented languages like Java there is no standardized way to calculate those metrics [7], therefore we implemented them to the best of our knowledge. Furthermore rule checks of Sonar can be used to verify the adherence to coding standards and best practices.

To support the verification of project specific requirements our implementation provides the possibility to calculate a measure using a custom defined formula that can make use of other measures and violation counts. The evaluation of this formula is done utilizing the Math Expression Parser of the Symja project[7]. For Software Escrow we propose *CCC*, a metric that sets Cyclomatic Complexity (CC) and comment lines density (C) in relation. Cyclomatic Complexity indicates the effort of an external developer to understand source code, documentation tends to ease understandability:

$$CCC = CC/(1 + (C/100))$$

Usage of third party resources. We further extended the framework by some digital preservation concerns that are also useful for escrow such as ensuring the availability of external sources. References to external third party resources, like libraries or Web services used in the software, can affect the functionality of the software. If the provider of the service is not available anymore, this can lead to a non-functioning program. Therefore external references have to be identified and properly inspected when verifying the source code. It has to be ensured that the service they are using is available in the long term. A potential strategy of the escrow agent is to deposit the library or materials needed for a Web service as well. If this is not possible, e.g., in the case of proprietary Web services, it has to be ensured that the executing source code sections are identified and reported as a risk to the customer. The use of external services should be specified in the licenses and escrow contract. Our implementation supports the escrow agent in identifying those external resources by reporting matches of a text-based *regex* search over source files which looks for Web service calls, system calls, etc.

[5]https://github.com/SonarSource/sslr
[6]http://textcat.sourceforge.net
[7]https://code.google.com/p/symja

A scenario that the escrow agent needs to be aware of is potential hiding of functionality in compiled libraries that limit the possibility to maintain the software. Instead of providing the source code, developers could supply compiled libraries to hide implementation details. Unknown libraries or those that are not available in public repositories are potential candidates for hiding code. To verify libraries our implementation performs a hash based lookup in the Maven Central Repository[8] of the JAR files that are part of the software. Other artifacts are looked up in the National Software Reference Library[9], a database containing hash values and other metadata of files that are part of software packages like Adobe Photoshop, Red Hat Linux, etc. Libraries which are not found in the corresponding database are reported and need to be checked against the agreements specified in the contract.

Legal certainty. Licenses are essential as they specify the legal foundation for the usage of the software. Thus the escrow agent needs to determine the licenses of the software's artifacts, as it has effects on the allowed usage in case of a release of the material. Our implementation extracts and identifies license information embedded in source files using the Perl script *licensecheck*[10]. For the licenses of the included libraries we use the License Maven Plugin[11] to determine the license information.

3.3 Redeployment Phase

The redeployment phase is the third phase of Software Escrow. Its main task is to ensure the quick release of the deposited material once a contracted trigger event occurs. The objective is to prevent a potential downtime of the customer's software. The events leading to the release of the software were agreed on in the *Planning* phase (cf. Section 3.1). If one of them occurs, the customer has to inform the escrow agent, who needs to check the contractual correctness, and verify the event. The agent is then obliged to release the material to the customer. Trigger events that lead to the release of the deposited materials to the customer can be the insolvency of the software developer, the liquidation of the developer's company, or an unjustified refusal of the developer to maintain the software.

4. EVALUATION

As an exemplary use case the review of aTunes[12], an open source audio player and media library, is tested. The case study performs a verification by using the criteria described in Section 3.2, similar to those done by an escrow agent. Figure 2 shows the Sonar overview presentation of the results from the technical framework, presenting the metrics and checks of the software. aTunes 3.0.8 consists of 81,915 lines of code and 1,499 classes. As material to deposit we used the sources in the SCM repository[13]. As software binary release that has been handed over to the customer we used the official release package[14].

Completeness of artifacts. The check for completeness of artifacts was done by building the software. All essential artifacts needed for the development of the software were contained in the deposit. Some documentation like requirements documentation, coding guidelines, and user documentation was available in the aTunes Wiki[15]. Other documentation for developers, like architecture description, was not found by manual inspection. Depending on the requirements, all documentation should be available in the deposited materials.

Consistency of sources and released binary. In order to compare binaries it is important to use the same compiler version for all builds. Otherwise the resulting binaries can differ even if the same sources have been used. After using the correct compiler and ignoring files that hold metadata like build count, build time, etc., there are still some files missing. Those are related to builds for other platforms than Linux, which we did not execute, and are not required when running aTunes in Linux. Besides that, the rebuilded artifacts match the reference artifacts.

Quality of documentation. The documentation of aTunes only consists of the comments in the source files, thus only these were evaluated. The 18.7% comment line density nearly matches the average of comment line density found in open source projects as determined in [1]. Further inspection showed that there are two abstract classes and 17 other classes with public methods that do not have any documentation. Interfaces and enumerations seem to be commented the most, which is good common coding practice. From the 1,441 source files in aTunes, the framework reported four to contain comments in Hungarian when expecting comments in English only. Manual inspection showed that all of the reported files actually contain comments written in English, so those were false positives. Other comment quality issues have been identified by our framework using the text proof tool LanguageTool. To reduce issues with little impact, like warnings about duplicate whitespaces, a reduced set of LanguageTool categories has been used to check for readability issues. The framework reported 133 text proof issues, which indicates a good overall quality compared to the length of the documentation.

Quality of source code. aTunes showed a Cyclomatic Complexity of 1.6 per method, 9.0 per class, and 9.4 per file. We compared the complexity of aTunes to the average complexity of the projects listed in the public Sonar instance Nemo [22]. At the time of evaluation Nemo contained 204 projects, 177 of them were Java projects, amongst others in-

[8] http://search.maven.org
[9] http://www.nsrl.nist.gov
[10] http://www.beathovn.de/licensecheck
[11] http://mojo.codehaus.org/license-maven-plugin
[12] http://www.atunes.org
[13] http://sourceforge.net/p/atunes/code/HEAD/tree/tags/aTunes 3.0.8
[14] http://sourceforge.net/projects/atunes/files/atunes/aTunes 3.0.8
[15] http://www.atunes.org/wiki/

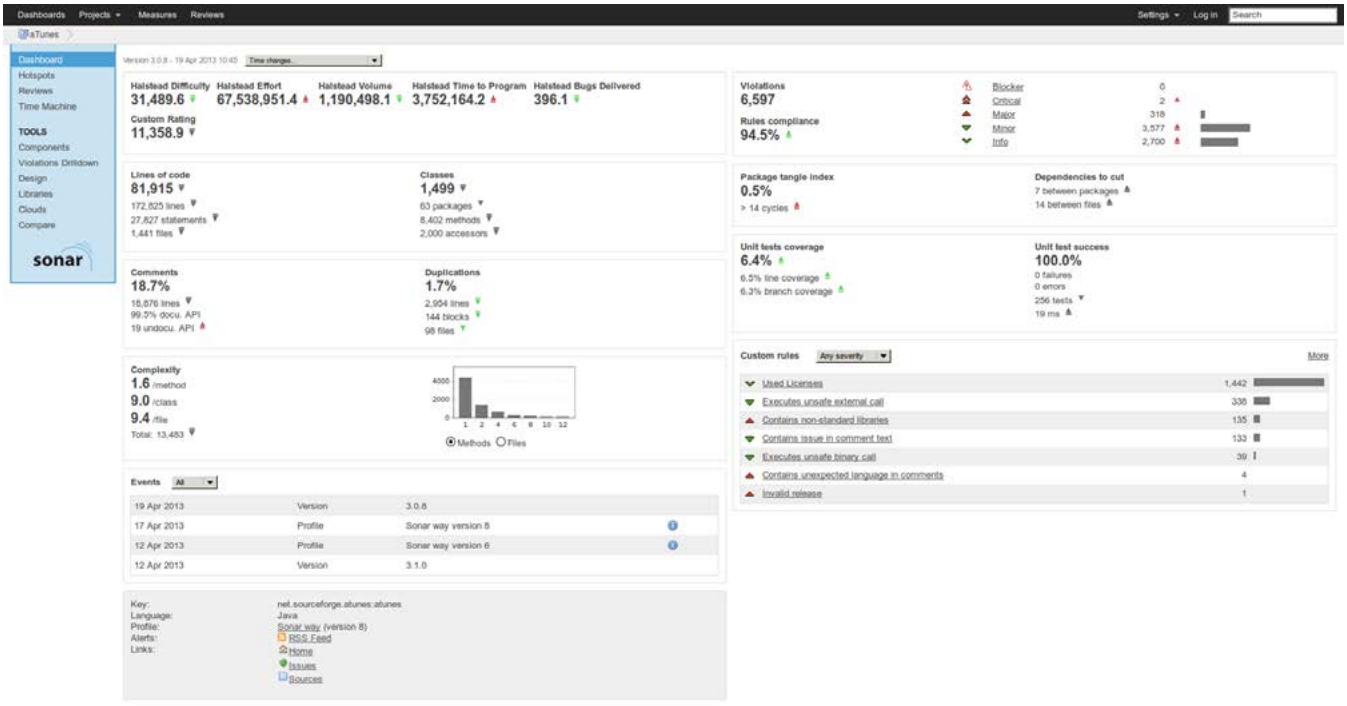


Figure 2: The extended Sonar dashboard showing aTunes' results

cluding the OpenJDK 7[16], JFreeChart[17], and several projects from the Apache Software Foundation[18]. The average Cyclomatic Complexity of those projects was 2.5 per method, 16.2 per class, and 19.5 per file. So considering other open source projects the results of this metric indicate a low complexity of aTunes. The Cyclomatic Complexity and Halstead's Difficulty metric of aTunes show similar results.

In our experiments we used a custom measure that takes into account the complexity and the comment density, with the idea that complex code should be easier to understand if it is commented properly (see Section 3.2). A list of the worst performing files gives a good starting point for a manual inspection of the software by the reviewer. Figure 3 shows the resulting poorest performing classes of this metric.

DeviceHandler	115.6
PlayListHandler	101.1
AudioFile	92.0
RepositoryHandler	87.6
DefaultTag	83.3
LastFmCache	80.1

Figure 3: The most incomprehensible classes of aTunes according to the CCC metric

[16]http://openjdk.java.net/
[17]http://www.jfree.org/jfreechart/
[18]http://apache.org/

Usage of third party resources. aTunes fetches most of its libraries through the build and dependency management tool Maven. Those libraries are considered trusted as they are provided from a central, public repository, which also provides the library's sources if available. Maven allows including additional repositories, which should be examined by the reviewer to verify their trustworthiness. From the nine libraries that are not obtained using Maven but already included in the deposited material, five have been reported as unknown, due to the fact that they were not available through the Maven repositories or that their hash value did not match one library there. All of them are used to create installation routines. The further investigation of the libraries depends on the agreements made in the escrow contract.

In our experiments our framework brought up 88 files that are assumed to contain external calls, 71 of them Web service calls and 17 binary calls. The number of Web service calls can be explained by further examination, which showed that the source of the calls are for instance modules that fetch additional meta data about the media from services like *Last.fm*. As we considered these modules optional, none of the Web service calls are an issue in this case. For a full functionality though, these Web service calls would pose a problem as they cannot be deposited as well. Inspection of the reported files showed further that many times operating system processes are spawned in order to execute external tools. One example of such a binary is *mplayer*[19], one of the supported audio playback engines. Other externally executed tools handle importing audio CDs to aTunes and encoding different audio formats. The deposited material

[19]http://www.mplayerhq.hu

contains binaries of the external tools for Windows and Mac OS. In Linux aTunes expects those tools to be installed in order to use the full functionality of the program. As mentioned in Section 3.2, externally called binaries are difficult to maintain so the reviewer should deposit those dependencies which are a necessity.

Legal certainty. All source files of aTunes contain license information (GPL in this case), but the framework could not find licenses in the JAR files that are part of the source distribution. In six cases the JAR files contained no licenses and in three cases the format of the license text could not be handled due to formatting issues. The licensing of those libraries need to be clarified in order to avoid legal consequences when releasing and further developing the software.

Summary. The applicability of the framework was shown in this Section. The framework supported review of aTunes showed that all artifacts required to build the software are available. External runtime dependencies are provided as artifacts, except those for Linux environments, which were missing. As discussed this could lead to problems in the future and the missing binaries should be put into escrow. Dependencies to Web services are used for optional features of the software. Depending on the requirements, these need to be put into escrow as well to preserve the full functionality of the software. The review also indicated that core parts of the software are well documented in general, which helps developers to familiarize with the software in a fine grained level. The lack of documentation of the architecture slows down understanding the big picture of the software design. Comparison to other software projects indicates that aTunes is not overly complex. There are no severe legal issues to be expected, as licensing of the artifacts is clearly specified with the exception of some installer tools that can be replaced without endangering the functionality of the software.

5. CONCLUSIONS

Software Escrow is a mitigation strategy when using a software developed by a third party. This paper aimed at presenting the necessary aspects needed for a successful Software Escrow, pointing out shortcomings of current practice, and presenting legal and technical considerations of this process. We also looked into the three different phases of Software Escrow, beginning with planning and setting up an agreement, executing the escrow by depositing the escrow material and verifying its quality with regard to maintainability, and finally redeploying the software. We further extended escrow by some Digital Preservation aspects, such as the use of external services that can be unavailable in the future. For the execution phase and its verification part we developed a Technical Software Escrow Framework by extending Sonar, an open source Software Quality tool, with escrow specific checks. This framework is able to check all kinds of material necessary for a successful deposit, from licenses over source code to documentation. By highlighting and reporting artifacts that have low quality it is able to support the verification of requirements agreed on in the escrow contract. We applied our framework for demonstration purposes to the open source software aTunes and analyzed the performance of our tool. It can be shown that Software Escrow critical parts of the software are found and reported back. These reports can then be used to easily find potentially problematic sections that need further improvement. Our framework thus achieves the objective to support a reviewer in analyzing the deposited material by partly automating the search for common software project issues.

6. ACKNOWLEDGMENTS

This work has been co-funded by COMET K1, FFG - Austrian Research Promotion Agency and by the TIMBUS project, co-funded by the European Union under the 7th Framework Programme for research and technological development and demonstration activities (FP7/2007-2013) under grant agreement no. 269940. The authors are solely responsible for the content of this paper.

7. REFERENCES

[1] O. Arafat and D. Riehle. The commenting practice of open source. In *Proceedings of the 24th ACM SIGPLAN conference companion on Object oriented programming systems languages and applications*, OOPSLA '09, pages 857–864, New York, NY, USA, 2009. ACM.

[2] C. Arapidis. *Sonar Code Quality Testing Essentials.* Community experience distilled. Packt Publishing, Limited, 2012.

[3] W. B. Cavnar and J. M. Trenkle. N-Gram-Based Text Categorization. In *In Proceedings of SDAIR-94, 3rd Annual Symposium on Document Analysis and Information Retrieval*, pages 161–175, 1994.

[4] CEN Workshop Agreement. *ESCROWGUIDE - Source Code Escrow - Guidelines for Acquirers, Developers, Escrow Agents and Quality Assessors.* European Committee for Standardization (CEN), 1999.

[5] CEN Workshop Agreement. *ESCROWGUIDE - Source Code Escrow - Guidelines for Acquirers, Developers, Escrow Agents and Quality Assessors - Part 3: A developer's guide to taking part in source code escrow.* European Committee for Standardization (CEN), 1999.

[6] CEN Workshop Agreement. *ESCROWGUIDE - Source Code Escrow - Guidelines for Acquirers, Developers, Escrow Agents and Quality Assessors - Part 4: A guide to providing a reliable escrow service.* European Committee for Standardization (CEN), 1999.

[7] D. De Silva, N. Kodagoda, and H. Perera. Applicability of three complexity metrics. In *Advances in ICT for Emerging Regions (ICTer), 2012 International Conference on*, pages 82–88, 2012.

[8] D. Draws, S. Euteneuer, D. Simon, and F. Simon. Short term preservation for software industry. In *Proceedings of the 8th International Conference on Preservation of Digital Objects (iPres 2011)*, pages 130–139, 2011.

[9] M. H. Halstead. *Elements of Software Science (Operating and programming systems series).* Elsevier Science Inc., New York, NY, USA, 1977.

[10] S. Helms and A. Cheng. Source Code Escrow: Are You Just Following the Herd?

http://www.cio.com/article/187450/Source_Code_Escrow_Are_You_Just_Following_the_Herd_?page=1&taxonomyId=3000, 2008. [Online; accessed 13-June-2013].

[11] G. Heyer, U. Quasthoff, and T. Wittig. *Text Mining: Wissensrohstoff Text*. W3L Verlag, Bochum, 2005.

[12] T. Hoeren, B. Kolany, S. Yankova, M. Hecheltjen, and K. Hobel. *Legal Aspects Of Digital Preservation*. Edward Eldgar Publishing, 2013.

[13] Iron Mountain Incorporated. Comprehensive Asset Verification and Testing. http://www.ironmountain.com/Services/Technology-Escrow-Services/Escrow-Verification-Services.aspx. [Online; accessed 17-June-2013].

[14] ISO 25010:2011. Systems and software engineering – Systems and software Quality Requirements and Evaluation (SQuaRE) – System and software quality models, 2011.

[15] ISO 9126:2001. International Standard ISO/IEC 9126, Part 1, Software engineering - Product quality - Quality model, 2001.

[16] M. Karger. Software-Hinterlegungsverträge. In *Computerrechts-Handbuch*. Kilian/Heussen, 2011.

[17] T. J. McCabe. A complexity measure. *IEEE Transactions on Software Engineering*, 2(4):308–320, December 1976.

[18] M. Miłkowski. Developing an open-source, rule-based proofreading tool. *Software - Practice & Experience*, 40(7):543–566, June 2010.

[19] NCC Group. Types of Verification. http://www.nccgroup.com/en/our-services/software-escrow-verification/software-verification/types-of-verification/. [Online; accessed 17-June-2013].

[20] M. R. Overly. *A Guide to IT Contracting: Checklists, Tools, and Techniques*. Auerbach Publications, Har/Cdr edition, Dec. 2012.

[21] V. Siegel. Software-Escrow. *Informatik-Spektrum*, 28(5):403–406, 2005.

[22] SONARSOURCE SA. Nemo - Sonar. http://nemo.sonarsource.org/, 2013. [Online; accessed 12-April-2013].

Leveraging DP in Commercial Contexts through ERM

José Barateiro
National Laboratory for Civil Engineering – LNEC
Av. Brasil, 101
1700-066 Lisbon, Portugal
jbarateiro@lnec.pt

Daniel Burda
SAP Research
Althardstrasse 80
8105 Regensdorf, Switzerland
daniel.burda@sap.com

Daniel Simon
SQS AG
Stollwerckstraße 11
D-51149 Cologne, Germany
daniel.simon@sqs.com

ABSTRACT
Until now, digital preservation research has been mainly driven by public or publicly funded organisations. The justification of costs for the preservation is based on abstract risks such as the risk of losing cultural heritage information, or the risk of data deficiencies for current and future research in big sets of data. Typically, the benefits from digitally preserving the objects of interest is difficult or impossible to quantify in terms of return-on-invest. In fact, it is common that memory institutions are mandated to preserve specific digital objects, making digital preservation not an option, but a legal obligation. While in the case of cultural heritage and scientific research qualitative reasons for preservation suffice, enterprises have an additional obligation to quantify the expected benefits and expenses in order to determine the scope of information to be managed and take commercial decisions for or against digital preservation. To provide appropriate means for leveraging the benefits of digital preservation in a commercial context, we argue in this paper that enterprise risk managers are the established function to assess and support decisions about preservation in enterprises. We show that enterprise risk management can be linked to digital preservation and how intelligent enterprise risk management can be utilised to identify the need for digital preservation, determine the corresponding actions, and contribute to the overall commercial success of enterprises.

Categories and Subject Descriptors
H.3.7 [**Digital Libraries**]

General Terms
Management, Measurement, Design.

Keywords
Digital Preservation, Intelligent Enterprise Risk Management, Commercial Use of digital preservation

1. OVERVIEW
The ubiquity of information technology in today's economies results in society's dependency on vital business processes supported and enabled by information technology systems. A vast amount of business, scientific and cultural information assets are created, filed and accessed digitally today. This digital information is a fundamental element for business success.

Society's dependency on digital processes conduces to a high exposure to risks affecting the businesses and the underpinning IT infrastructure. Continued access to digital data cannot be taken for granted [1]. Indeed, any business that deals with information can be subject to several risks that should be actively mitigated by digital preservation (DP) means.

DP can be understood as "the ability to sustain the accessibility, understandability and usability of digital objects ..." [2]. It ensures long-term access to digital information. The meaning of long-term has been defined in the OAIS standard (ISO 14721) as "long enough to be concerned with the impacts of changing technologies, including support for new media and data formats, or with a changing user community. Long Term may extend indefinitely". Accounting for this definition and considering the rapid development in information technology, the challenges of preserving digital information in its notion of an intangible asset becomes more and more pressing [3].

In commercial environments, many businesses are primarily focused on short term returns rather than long term sustainability. If DP methods can also be conceived as a means to mitigate business risks then DP can play an integral role in a commercial context. This paper describes how the European funded project TIMBUS [4] is addressing DP as a risk management activity in enterprise contexts.

The rest of the paper is structured as follows. In Section 2 we briefly describe the state of the industry with regards to Risk Management (RM), Enterprise Risk Management (ERM) and Intelligent Enterprise Risk Management (IERM). Section 3 explains how established RM can be extended to integrate DP. Section 4 explains potential benefits of DP for enterprises. In Section 5 we briefly summarise this paper and provide an outlook into future work.

2. RISK MANAGEMENT IN ENTERPRISES
Enterprises apply RM in their various business fields and have developed sophisticated risk assessment and evaluation methods for business domains such as financial, credit and market risks. While the specific risks vary and are heavily subject to expert knowledge, RM processes and methods have undergone standardisation. In the following, we use the generic ISO 31000 RM standard [5]. It formulates RM as an on-going process embedded in an organisational context. This standard has proven its applicability in our research project TIMBUS [4] and serves as the foundation for integrating RM and DP.

2.1 ISO 31000 overview
The ISO 31000 RM standard defines the principles and implementation of RM to control the behaviour of an organization with regard to risk. It is based on the principle that RM is a process operating at different levels, as shown in Figure 1. The RM process is characterized by the combination of policies and procedures applied to the activities of establishing the context; assessing (identifying, analysing and evaluating); treating; communicating and consulting; and monitoring and reviewing the risks.

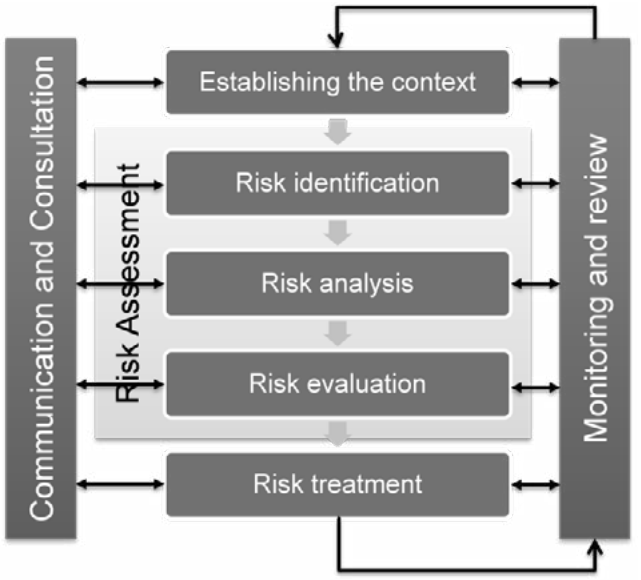


Figure 1. RM process according to ISO 31000

First, establishing the RM context is crucial to identify strategic objectives and define criterions to determine which consequences are acceptable to this specific context. Second, today's organizations are continuously exposed to several threats and vulnerabilities that may affect their normal behaviour. The identification recognizes the existence of risks; analysis examines the nature and severity of the identified risks; and evaluation compares the severity of risks with the defined risk criterions, to decide if the risks are acceptable, tolerable or define the appropriate techniques/controls to handle them.

The identification of threats, vulnerabilities and risks is based on events that may affect the achievement of goals identified in the establishing the RM context phase. Different methods such as brainstorming, questionnaires or inspection support identification of risks. Whatever method used, it is crucial to be as open minded and holistic as possible, because any risk not identified in this step cannot be evaluated in the following steps. For simplifying the understanding and handling, risk managers create taxonomies for risk sources as well as for impact areas. These taxonomies offer the possibility to aggregate the risks to a higher level enabling the required level of abstraction for an effective and efficient RM. To achieve maximum accuracy and completeness, it is best practice to use systematic approaches as offered by Quality Risk Management (QRM) [6] for initialising the identification of risks.

After risk identification, the risk analysis and evaluation estimates the likelihood and impact of risks to the strategic goals as to be able to decide on the appropriate techniques to handle these risks (risk treatment). To determine the likelihood of events and their consequences, probabilities can be estimated and underpinned with indicators. Since the level of risks depends on the effectiveness and efficiency of controls in place existing controls are assessed for their practical relevance to the respective risks. Risk treatment options include:

- avoiding the risk by deciding not to start or continue with the activity that gives rise to the risk;
- taking or increasing risk in order to pursue an opportunity;
- removing the risk source;
- changing (for negative impacts reducing) the likelihood;
- changing the consequences;
- sharing the risk with another party or parties (including contracts and risk financing); and
- retaining the risk by informed choice.

The risk treatment step executes per risk the treatment as determined before to reduce the risk and/or to mitigate it. The RM process requires a continuous monitor and review activity to audit the behaviour of the whole environment allowing, the identification of changes in risks, or the suitability of implemented risk treatment procedures and activities. Finally, the communication and consultation activities are crucial to engage and dialog with stakeholders.

2.2 Application in industry

Many industries implement business changes through projects. Several reference frameworks for project management are established to give guidance when initialising, operating and finalising a project. The most widely-known reference frameworks are PRINCE2 [7], a structured project management framework from the Office of Government Commerce in UK; Project Management Body of Knowledge (PMBoK) [8], a reference to body of knowledge for project management from the Project Management Institute in USA; and IPMA Competence Baseline (ICB) from the International Project Management Association [9].

In project management, a risk is defined as a possible event or circumstance that can have adverse influences on the outcome of a project. RM manages these events, their negative impacts and initiates mitigation actions accordingly. All of the above frameworks cover RM as an integral part. Note that RM does not directly affect or improve project outcomes (e.g., deliverables or work products), but gives additional insights into and transparency about the project outcomes' status and allows for mitigation actions to influence the future course of actions.

While RM is often used in an isolated way (e.g. per business area or per country), ERM breaks the thinking in silos and establishes a holistic enterprise wide management of risks. To address risks at the organisational level and integrating the different views of the stakeholders, ERM provides a framework to manage the uncertainty and the associated threats and opportunities in the context of an enterprise. An example for an integrated model with a strong history in financial auditing is the COSO Enterprise RM framework [10].

The Accenture Global Study [11] reveals a growing importance of ERM. More than 80% of survey respondents have an ERM program in place or plan to have one in the next two years with European companies being the least likely to have an ERM programme (at only 52%). Many companies have started to appoint C-level oversight of the RM function or even establish Chief Risk Officers. The study reveals that 83% of executives expect their investments in RM to increase over the next two years. The bottom line: there is a strongly growing market in RM capabilities and we should aim at triggering DP via RM.

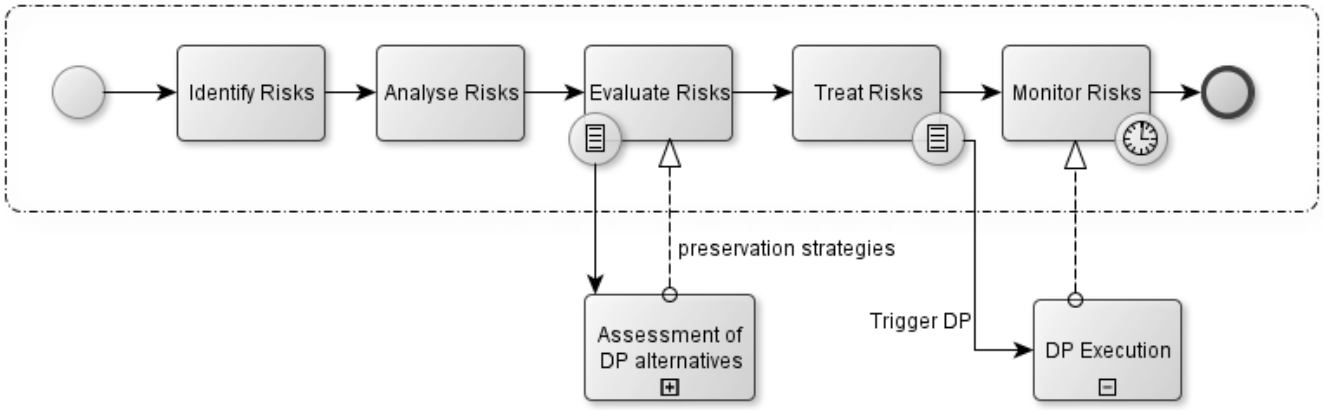


Figure 2. Integration of RM into DP of business processes

2.3 Risk management in digital preservation

The DP community has considered and integrated RM concepts to assess DP repositories. The TRAC Criteria and Checklist [12] is meant to identify potential risks to digital content held in repositories.

The Digital Repository Audit Method Based on Risk Assessment (DRAMBORA, cf. [13]) process focuses on risks, their classification and evaluation according to the activities, assets and contextual constraints of individual repositories. It aims at traditional DP scenarios, providing a catalogue of typical risks in DP environments. In this paper, we take a different point of view with regards to RM and try to elaborate how DP can be beneficial in scenarios where DP is not required *per se*, for example, in enterprises where running business processes are used for commercial purposes [14], [15] or e-Science [16] where the information must be permanently available.

3. DIGITAL PRESERVATION AND RISK MANAGEMENT

DP contributes new aspects to the overall process of ERM in terms of risk identification, risk analysis, and risk evaluation. For certain risks, DP provides effective and efficient means of risk mitigation, i.e., it either eliminates the sources of the risks (e.g., data loss) or at least reduces the likelihood or negative consequences of risks (e.g., availability risks due to failure of disaster recovery).

The goal of DP is the preservation of and within RM, DP provides a toolset for handling and mitigating information related risks. The establishment of interfaces between DP and RM is therefore essential. As shown in Figure 2, the RM process is composed by: *Identify Risks, Analyse Risks, Evaluate Risks, Treat Risks and Monitor Risks*. The *Assessment of DP alternatives* is an external activity used by *Evaluate Risks* to evaluate DP solutions against any other potential RM strategies. Finally, if DP is selected as the treatment action for a particular business risk, the risk treatment process will trigger the archive to start the necessary preservation activities through the *DP execution* process.

3.1 Identify risks

The identification of risks in the organizational context can be extended by identifying risks related to information obsolescence (the original motivation for DP in other areas). In particular, we have identified the following risk areas to be relevant for enterprises:

- Compliance risks;
- Audit risks;
- Business Continuity risks;
- Legal risks, in particular intellectual property (IP) rights;
- Operational risks; and
- Competition risks.

3.2 Analyse risks

After risks have been identified they have to be analysed along the dimensions of their impact and probability of occurrence to obtain an adequate loss estimate representing the financial impact for the organization. We propose a dependency model to systematically investigate interrelations between risks and business processes including their underlying process activities and supportive IT components such as hardware and software.

- Business processes in an organization consist of a defined set of process activities geared towards the efficient execution of a business process.
- The process activities, in turn, are increasingly supported by both internal and external IT components and services.
- The IT components are exposed to different risks, which can result in their unavailability. As a consequence, process activities that are supported or realized by these IT applications cannot be executed. The unavailability or failure an activity has, in turn, an adverse effect on the business process since certain process activities cannot be executed.

To uncover which business processes are affected by which risks, the dependency model can be formalized by the means of three matrices that represent the various layers and their inherent dependencies (cf. [17]):

- Relationship between business processes (BP) and process activities (A) BP×A;
- Relationship between process activities and IT components (ITC) A×ITC; and
- Relationship between ITCs and risks (R): ITC×R.

The first matrix (BP×A) describes the relationships between business processes and its constituting process activities represented by the probabilities of an activity being executed during business process execution. These probabilities can be obtained by means of the business process model. Elements in the matrix can take

values between 0 (activity not part of business process) and 1 (activity part of business process and always executed during business process execution).

The second matrix (A×ITC) reflects which process activities are dependent on particular IT applications. Thirdly, matrix ITC×R represents which IT components are affected by which risks. The relationships are modelled in a binary manner whereby "1" means that a risk affects a specific IT component and "0" means it does not affect it.

To ultimately derive the cause-effect relationships between risks and business processes the three matrices described above have to be multiplied through all layers leading to the matrix

$$BP \times R = (BP \times A) * (A \times ITC) * (ITC \times R).$$

Once the relationships have been identified, it becomes obvious which business processes are affected by which risks and to what degree according to the flow of process activities. Based on those findings risk managers are able to proceed with an adequate determination of quantitative loss values for the organization, reflecting the financial impact for an organization. In an effort to calculate the expected cost of a risk, a widely accepted approach is to build the product of a risks' likelihood and impact level [18]. Determining a risks' likelihood is one of the most challenging parts of qualitative risk analysis since often little historical data are available. In that case, external risk databases can be used to support the determination of the likelihood.

Besides the financial dimension of a risk extant research provides suggestions on additional components of impact attributes that help to better determine the overall risk level [19] as quantitative methods lack the ability to provide a holistic analysis of secondary impacts [20]. Secondary risk impact values are not measured in numeric terms but rather as verbal, discrete statements [21]. Towards the end of a holistic approach that not only considers the direct financial impact caused by a risk but also considering secondary impacts we draw from [22] and suggest a framework of secondary impact attributes to include the dimensions of strategic, reputational, customer and legal impact.

3.3 Evaluate risks

The next step in the process is the assignment of risk classes and the comparison of different risks. For each of the risks identified before, the risk manager determines mitigation actions for risks, i.e., for risks where DP can be used as a mitigation action, he considers DP as risk treatment. As to decide whether DP is a suitable treatment, the following criteria are taken into account:

- cost of DP in different service levels;
- value at risk in business process,
- underlying activities, and supporting IT; and
- residual risk with digitally preserved business process.

3.4 Treat risks with digital preservation

In the area of information related risks, DP can assist at the following three aspects of risk treatment:

- Changing the likelihood of specific risks. Establishing DP is expected to lead to more transparency about business processes in organizations. Many of the risks addressed by DP are caused by informational lack of transparency.
- Change the (negative) consequences of adverse events, e.g., facilitating disaster recovery, enabling business continuity
- Sharing the risk with another party or parties: DP assures availability of information. In this respect, DP will move the information related risk to archive providers who will have to deal with archive related risks.

A full and detailed catalogue of enterprise risks where DP affects has to be developed in the specific context of an enterprise. In general, DP does not focus on domain specific business risks such as credit risk, counterparty risks, currency exchange, etc. but mainly treats information related risks. Since information is derived from data relative to specific contexts, risk identification and DP need to be tailored to the environment as required. Amongst others, DP affects

- Compliance risks;
- Audit risks;
- Business Continuity Management (BCM) risks;
- Legal risks;
- Operational risks; and
- Competition risks

as will be discussed in Section 4.

3.5 Monitor risks

For risks where DP is a feasible treatment, often actions needs to be taken due to changes of technology or the context. In general all changes of a processes context may lead to information related risks. Examples for changing contexts are

- Organisational changes (service providers go out of business or are acquired by a different company); or
- Legal changes (regulation, taxation, IP rights).

From the RM perspective, the Risk monitoring provides the DP governance and management layer in terms of the business. DP planning is triggered when the Risk Manager identifies the need for DP as a mitigation action (or, more general, as a risk treatment). The Risk Manager is responsible for providing a rough business case. After DP planning, the rough cost estimate is validated against the business case and DP design and DP execution are completed.

In case the risk events occur and have the anticipated (negative) impact, the monitoring and control process triggers the DP access step. To this end, any of the risk events as identified in risk analysis can trigger the DP access according to the risk mitigation plan. Additionally, the DP internal monitoring and control process needs to be established to maintain the structure of the digitally preserved business process vitality.

3.6 Roles and responsibilities in RM

To perform the RM process steps described above in an accurate manner that is aligned with an organizational objectives, specific roles and responsibilities need to be defined and assigned within the organization. Therefore, RACI charts have proven to be useful means in the project management arena. A RACI matrix describes the participation by various roles in completing specific activities. Extrapolating to the context of this study, RACI matrices can support the clarification of roles and responsibilities required to perform the RM processes and related DP activities. Extant research indicates that the organizational configuration of DP activities in a corporate context is contingent on internal and external factors. Thus, we propose to employ RACI matrices in support of RM and DP to appropriately assign responsibilities as illustrated in Table 1.

R: Responsible A: Accountable C: Consulted I: Informed	Organizations Management	IERM Manager	DP Manager	Indicator Manager
Risk Evaluation				
Calculate and assess risks	C	RA	C	
Determine risk treatments	C	A	R	
Generate reports		A	R	C

Table 1. Illustration of a RACI matrix

4. DIGITAL PRESERVATION PROCESS BENEFITS

A traditional cost/benefit analysis is an approach to measure benefits and costs. Although, costs for a DP program often do not directly map to costs in other programs, making it extremely difficult for decision makers to create an accurate budget for preservation. In the following, several use cases are elaborated and the respective benefits for the stakeholders of the use case are qualified. The success and acceptance of DP in industry can be fostered if ERM identifies benefits and the specific risks to be mitigated by DP. These benefits can be pinpointed at least to the following use-cases.

4.1 Compliance and Regulatory Requirements

In almost all industries and markets, authorities define rules and regulations for the market players either because the markets are of highest importance for European Society as a whole or the markets are dominated by a small number of big players and the European Monopoly Commission monitors the market behaviour to assure fair pricing for end consumer. Examples for regulated industries and markets are amongst others telecommunications, energy, and banking sectors and the respective markets. To demonstrate market behaviour according to the rules and regulations becomes more and more complex but is ever more closely monitored by authorities and auditors. According to [11], one of the biggest challenges in RM is the implementation of regulatory demands and the compliance with the rules and regulations is the most business critical driver for future activities.

4.2 Transparency on Intellectual Property Rights

In today's commercial environments, business processes and the supporting IT environment demands for proper management of IP rights. Numerous artefacts of different types are utilized and made use of to achieve the overall business objectives. With DP, all relevant artefacts and artefact types are identified during the archiving process, e.g.,

- Services (subscription licenses);
- Software (license keys for applications);
- Databases (licenses for DBMS); and
- Content (videos, pictures, music, text, …).

The different types of artefacts usually come with different types of IP rights. In everyday use, to make use of an artefact protected by IP rights a license from the owner of the respective right needs to be acquired by the user of the artefacts. Even though license management is a standard task in IT Service Management, many companies have room for improvement in the day to day implementation. As different countries have different regulations concerning the treatment of intellectual property right, there is a significant risk that IP rights are violated in daily business and business processes depend on proper licensing. In some cases, companies have been sentenced to pay enormous amounts of license fees to the IP owners. Additional complexity comes from diversification of the IP rights depending on the artefact type.

As part of DP, IP rights for the various artefacts are identified during expediency and tested during exhumation. If exhumation is tested properly (e.g. into an environment sufficiently different from the origin environment), IP gaps such as missing licenses can be detected and fed back to the license management functions.

A second aspect comes into play when the originator of business process, software, or other work wants to prove authorship of certain artefacts. In this case, DP can be used to provide evidence of the state of the art at the time of DP execution. (If a 3rd party intends to open a case for patent rights about a process, software etc. the evidence of 'prior art' can be made by disclosing the DP archive and make use of the archive provider as a 'neutral' witness).

4.3 Long-term Customer Support

Certain industries (like airplane or pharmacy industries) sell products with long lifecycles or the products are based on a rapidly changing technical platform. If a company wants to provide long-term support to their customers either for the products it is worth considering DP as an enabler for long-term preservation of business and product related side products, processes and knowledge.

In IT focussed organisations, often IT service management frameworks (ITSM), in particular, IT infrastructure library (ITIL) [23] as best practise approach is applied to ensure the quality of support. The service operation processes of ITIL as well as a similar process structure in non-IT organisations can be regarded as the set of business processes delivering support for the customer. If an organisation applies the concepts and methods of TIMBUS DP to this set of business processes, long term support for customers can be achieved. In an ITIL based organisation environment, a number of concepts from ITIL (e.g., the Definitive Media Library (DML) where all configuration items including associated items like documentation and licenses) can be re-used in the DP context. DP assures availability and accessibility of significant and relevant information to that even after a long period of time all knowledge required to support a product or a service is retained and preserved even after the service itself has been decommissioned and can be recovered easily.

4.4 Competitive Advantage

Competitive advantage is achieved when an organisation adopts or develops a capability or combination of capabilities that allows it to outperform its competitors. With DP in place, commercial organisations have a number of competitive advantages over other market players with DP.

Firstly, an enterprise that is DP ready has achieved a maturity level that can be actively advertised to its clients. The enterprise has proven capabilities of pro-active and sustainable business process management and can demonstrate to clients its modularisation and standardisation of business processes. In other words, DP ready organisations are well advanced on their path to an industrialised IT and have repeatable and predictable processes. As a consequence of the process oriented work, the enterprise can leverage the benefits of division of labour and make use of out-

sourcing methods to lower costs on one hand. On the other hand, due to internal resources focussing on their competencies the services and products can be evolved and enhanced much faster than in an everyone-does-all working style.

Secondly, in specific environments, DP readiness can be a distinctive feature – e.g., in the public sector, avionics, or defence industry as it shows the long term strategic approach of an organisation to the market.

4.5 Side effects of digital preservation

Establishing DP in organizations is expected to have positive side effects as well as negative ones. The first positive side effect is expected to be an increasing maturity of the organisation. Oriented on the different levels of the Capability Maturity Model Integration the improvement of organisations is correlated with the increasing degree of transparency. As DP needs a holistic transparent view on an organisation, the introduction of DP will automatically increase the maturity.

As DP is about the instantiation of the preserved environment in a new context, it is expected that DP will reduce the dependence of the artefacts to preserve from different persons. In this case DP will advance the enterprise on their way to an industrialised IT.

The increasing awareness for risks laid in information and business processes is expected to improve the awareness for the information and the business processes itself. If both are more present and exposed they can support the role of the business process management. The information about the objects to preserve and the contexts they are embedded will also lead to higher degree of transparency in the business process and the underlying (IT) artefacts.

On the other hand, DP may also lead to negative impacts. At least the instantiation of the archive will cause different efforts like every other entity in organizational processes. As DP is not directly affecting the core business it will lead to higher management efforts and increase administration overheads for the first time. Like every other monitoring activity, the maintenance of the digital archive (analyse the designated community) may slow down the daily business a bit.

An additional negative side effect could be the increased effort needed to address privacy policies within an archive. This will affect different administrative departments in an organization and increase the communication overhead.

5. SUMMARY AND OUTLOOK

In this paper we laid out the enterprise view on DP and how RM can be extended to be an advocate function for DP in commercial contexts. To this end, we propose to argue for DP as a risk treatment for certain business risks and show how DP processes can interact with established RM processes.

As a next step, the processes and concepts described above will be applied and evaluated in several use cases in the course of the TIMBUS project.

6. ACKNOWLEDGMENTS

Parts of this work have been supported by the European Union in the TIMBUS project [4]: “Digital Preservation for Timeless Business Processes and Services”, Grant Agreement Number 269940.

7. REFERENCES

[1] F. Berman, "Got Data? A Guide to Data Preservation in the Information Age," Communications of the ACM, vol. 51, no. 12, pp. 50-56, 2008.

[2] S Rabinovici-Cohen, M G Baker, R Cummings, S Fineberg, and J Marberg, "Towards SIRF: Self-Contained Information Retention Format," in *Proc. of the SYSTOR '11*, Haifa, 2011.

[3] CCSDS. (2011, Oct.) Reference Model for an Open Archival Information System (OAIS).

[4] TIMBUS. (2011-2014). http://timbusproject.net/about

[5] ISO, ISO 31000 Risk management — Principles and Guidelines, 2009.

[6] Frank Simon and Daniel Simon, *Qualitätsrisikomangement*. Berlin: Logos Verlag, 2010.

[7] Office of Government Commerce, *Managing Successful Projects with PRINCE2*., 2009.

[8] William R. Duncan, *A guide to the project management body of knowledge (PMBOK guide)*.: Project Management Institute, 2004..

[9] International Project Management Association, *IPMA Competence Baseline 3.0*., 2006.

[10] COSO. (2012, March) Committee of Sponsoring Organizations of the Treadway Commission. http://www.coso.org/erm-integratedframework.htm

[11] Accenture, "Report on the Accenture 2011 Global Risk Management Study," 2011.

[12] CRL/OCLC, "Trustworthy Repositories Audit & Certification: Criteria and Checklist (TRAC)," The Center for Research Libraries and Online Computer Library Center, 2007.

[13] A. McHugh, R. Ruusalepp, S. Ross, and H. Hofman, "The digital repository audit method based on risk assessment (DRAMBORA)," in *Digital Curation Center and Digital Preservation Europe*, 2007.

[14] J. Barateiro, G. Antunes, M. Cabral, J. Borbinha, and R. Rodrigues, "Digital preservation of scientific data," in *European Conference on Digital Libraries*, Aarhus, Denmark, 2008.

[15] J. Barateiro, "Digital preservation of heterogeneous data," Bulletin on IEEE Technical Committee on Digital Libraries 2009.

[16] D. Marcum and G. George, "The Data Deluge - Can Libraries Cope with e-Science," Libraries Unlimited, 2010.

[17] S. Sackmann, "A reference model for process-oriented it risk management," in *ECIS 2008 Proceedings*, 2008.

[18] B. Suh and I. Han, "The IS risk analysis based on a business model," *Information & Management* , vol. 41, no. 2, pp. 149-158, 2003.

[19] H. Beeck and T. Kaiser, "Quantifizierung von Operational Risk," in *Handbuch Risikomanagement*, L. Johannig and B. Rudolph, Eds., 2000, pp. 633-654.

[20] J. Hargreaves, "Quantitative Risk Assessment," in *Enterprise Risk Management*, J. Fraser and B. J. Simkins, Eds., 2010, pp. 219-236.

[21] H. P. Königs, *IT-Risiko-Management mit System*.: Vieweg + Teubner, 2005.

[22] J. Fraser and B. J. Simkins, *Enterprise Risk Management*.: John Wiley & Sons, 2010.

[23] Office of Government Commerce , *IT Infrastructure Library (ITIL)*.: The Stationery Office, 2007.

Benefits of geographical, organizational and collection factors in digital preservation cooperations: The experience of the Goportis consortium

Michelle Lindlar
German National Library of Science and Technology (TIB),
Welfengarten 1B, D-30167 Hannover
+49511/ 76219826
michelle.lindlar@tib.uni-hannover.de

Yvonne Friese
Leibniz Information Centre for Economics (ZBW)
Düsternbrooker Weg 120
D-24105 Kiel
+49431/8814610
y.friese@zbw.eu

Elisabeth Müller
German National Library of Medicine, Medicine. Health. Nutrition. Environment. Agriculture. (ZB MED),
Gleueler Straße 60, 50931 Köln
+49221/4785680
Elisabeth.Mueller@zbmed.de

Thomas Bähr
German National Library of Science and Technology (TIB)
Welfengarten 1B, D-30167 Hannover
+49511/76217281
thomas.baehr@tib.uni-hannover.de

Anja von Trosdorf
German National Library of Medicine, Medicine. Health. Nutrition. Environment. Agriculture. (ZB MED),
Gleueler Straße 60, 50931 Köln
+49221/4787078
Anja.Trosdorf@zbmed.de

ABSTRACT
Digital preservation is a resource intensive task, requiring specific systems, well-trained staff and an ongoing commitment to adopt new strategies and approaches as technology and/or user expectations change over the course of time. Cooperations to tackle this task are not a new idea - one of the first reports on digital preservation, commissioned by the Center for Preservation and Access (CPA) and the Research Library Group (RLG), recommended a "national system of digital archives" [12]. In the library context consortia date back to the 1970s, where they rose in the context of shared cataloguing efforts. Over the years experiences have been gained in different forms of cooperations and consortia. Some are focused on a grouping of institutions based on institution type or regional factors, while others are more cross-sectional services, focused around factors like material type. The Leibniz Library Network for Research Information (Goportis) consists of the three German National Subject Libraries [19]. Goportis conducted a digital preservation pilot project between the years 2009-2011 and is now operating a collaboratively used central digital preservation system. The paper highlights the lessons learned from experiences in the collaborative approach to digital preservation, focusing on the influence the factors "geographical location", "organization type" and "collection" have on a shared system implementation and operation. Based on a literature study of international best practices, guidelines and recommendations a thesis will be formulated for each of the three factors, which will then be checked against the experience gained by the Goportis consortium.

Categories and Subject Descriptors
H.3 [**INFORMATION STORAGE AND RETRIEVAL**]: Systems and Software – *Information Networks.*

General Terms
Management, Documentation, Reliability, Human Factors, Standardization, Legal Aspects

Keywords
Library Network, cooperative conducted Digital Archive, Consortial Digital Preservation

1. INTRODUCTION
Forming cooperations to tackle complex tasks is not a new phenomenon in the world of cultural heritage institutions. Collaboration, in contrast to a mere cooperation, refers to an "in-depth sharing and pooling of resources" [6]. The main motivation for engaging in library consortia is the benefit of sharing resources and experience. Collaborative cataloguing efforts and digitization projects date back to the 1970s. It thus comes as no surprise that the digital preservation world is now looking back at many years of experience in collaboratively run systems. Early initiatives in consortial digital preservation systems include system developments like DAITSS (Dark Archive In The Sunshine State), the preservation repository system of the Florida Center for Library Automation, targeted towards the 11 publically funded universities in Florida or the MetaArchive, an international collaboration and one of the first private LOCKSS networks in the world.

When looking for partners for collaboration, three factors usually play a role:

- geographical distance or association, e.g. in the form of city-wide or national cooperations
- organizational association, e.g. in the form of collaborations of library consortia or state archives
- collection factors (subject and/or material type based), e.g. in the form of collaborations to handle geospatial information or web-archiving

Goportis - the Leibniz Library Network for research information - consists of the three German national subject libraries: The German National Library of Science and Technology (TIB), the German National Library of Medicine (ZB MED) and the German

National Library of Economics (ZBW). The three Goportis partners have been conducting a digital preservation project since 2010, first an 18-months pilot and since the end of 2011 a running digital preservation system. It is our aim to support individual scientific workflows and research. We want to build a sustainable trustworthy digital preservation system for the three National Subject Libraries in Germany. Our three institutions have the mandate for Archiving and the responsibility for the long-term-access of our digital objects. Cooperative work supports learning from each other, bundling our resources and avoiding redundant work. Therefore, we can handle the task of digital preservation more cost-effectively and engage in efficient workflows.

The benefit of consortial digital preservation lies in working more effectively or efficiently, usually by bundling resources like staff skills and expertise. Furthermore, consortial digital preservation can be more cost-effective, if an out-of-the-box system is bought by the whole consortium or the storage is organized centrally (positive economies of scale). This definition meets the aim of the Goportis digital preservation collaboration, which, in contrast to a mere cooperation, refers to an "in-depth sharing and pooling of recourses" [6].

The chapters 2, 3 and 4 of this paper highlight the experiences gained by the three libraries by analysing the influence the factors "geographical location", "organization type" and "collection" have on a shared system in implementation and operation. Based on a literature review of international best practices, guidelines and recommendations, theses for each of the three factors are formulated and checked against the experience gained by the Goportis consortium. Within the scope of this paper, the analysis will focus on collaborations of different institutions in running a jointly operated digital preservation system. Collaborations with a mere focus on standardization or knowledge exchange, as well as collaborations where one partner only provides the development of the system, are out of scope.

2. Factor: Geographical location

A consortium of memory institutions, founded to cooperatively conduct a digital archive, can be based on geographical location, which means that all partners are located in the same country or even in the same state or county. As for implications, this may include national legislature or legal restrictions.

Public institutions like libraries, archives or museums, which are located in the same country or even in the same state, typically share other commonalities like the legislation and similar tasks and responsibilities. Digital preservation can be one of them. For German national libraries, for example, digital preservation is mandatory.

Institutions based in the same region are likely to belong to the same scientific community and to have already worked together before. Libraries which are located in the same part of Germany may be connected to the same union catalogue.

2.1 Analysis of existing guidelines, best practices and reported experience

The location of the collaboration is crucial [6] and geographical closeness makes collaboration "more likely to occur and easier when it happened"[7]. Besides, general cooperation benefits like sharing "physical resources such as space and conservation of collection" [7], carry more weight when the collaboration partners are located in the same region.

Collaboration based on geographical location bears specific benefits as well. In the UK, for instance, memory institutions like libraries and museums share the same policies, have to support the conservation of collective memory and share the same cultural identity [11].

In the following, three examples of geographically defined consortia in digital preservation are described. DAITTS - if you consider the FCLA service as well - is limited to the state of Florida; nestor and kopal are limited to Germany.

The long-term preservation repository service DAITTS, Dark Archive in the Sunshine State, was developed by the Florida Center for Library Automation (FCLA) and supported by the IMLS (Institute of Museum and Library Services). It is used by ten of the eleven publicly funded universities in Florida.

As all institutions are located in the same state, the same law is valid for all of them. When dealing with copyright issues, the same laws and rules are valid for all partners. Hence, findings should be shared to avoid redundant work and to use synergy possibilities of the cooperation.

Metadata standards often differ - sometimes even within the same institution and "different standards are often based on different formats"[4]. Interoperability is crucial when running a digital archive together.

The metadata schemas and standards used for the several catalogues in the University Libraries of Florida are heterogeneous, MARC21 is widely spread, but so are ALEPH, Dublin Core, AACR2 and many more. Although the university libraries are located in the same state, the commonalities do not extend to the metadata standards used, so a central solution or one workflow for all partners is not possible for reasons of heterogeneity. DAITTS is capable of dealing with the needs of the different consortium members, as it uses PREMIS and METS, so it is possible to embed several different kinds of standards in the metadata information of the Archival Package. Within DAITTS, the archiving institution is responsible for adding adequate descriptive metadata to its objects itself [5].

Nestor, the German competence network for digital preservation, was founded in 2006, and with 14 members it is the biggest consortium for digital preservation in Germany. The network aims to generate guidelines for topics related to digital preservation and to develop standardizations, e. g. for trusted repositories [22]. All partners contribute with their knowledge and experience and establish infrastructures (eight topic-based nestor working groups, public websites, internal wiki, newsletter via email, downloadable publications and guidelines) to share their findings and research transparently for either the other partners or even for the whole German community so that everybody can benefit. For instance, as memory institutions in Germany all share the same legislation, a new nestor working group was founded in 2012 to establish a guideline to develop preservation polices [22].

Kopal [21] started in 2004 and the partners were the National Library of Germany (DNB) [18], the State and University Library of Göttingen (SUB Göttingen), IBM Germany and the data centre GWDG (Gesellschaft für wissenschaftliche Datenverarbeitung in Göttingen). The project aimed to establish a cooperatively built and managed long-term preservation system for digital objects [1]. The system intended to implement the components of the OAIS model and to create the technical prerequisites for digital preservation. The storage solution is based on the DIAS system (developed by IBM), the ingest and access tools are based on the

open source software library koLibRI. The system was created in a way that several partners would be able to use it [2]. From the first, the metadata management of the DIAS system was configured generically, so that the needs of the different partners could be fulfilled [2].

The kopal test system was fully developed in 2010. Although the segmentation of the responsibilities between the project partners and the communication about technical standards and details proved to be complex, practical experience showed that it is functional and feasible [2].

Based on kopal, there was a follow-up-project, DP4Lib, funded by the German Research Foundation (DFG), which aimed to offer digital preservation as a service for public institutions in Germany [17].

2.2 Theses

2.2.1 Thesis 1: Being located in the same country simplifies the sharing of findings referring to legislation issues

A consortium based on geographical location bears an advantage referring to legislation. If all partners belong to the same country - or even to the same state - the same laws like e. g. the copyright law and telecommunications act apply to every partner. This makes it easier to stick to the same rules, to act upon the same policies and to organize issues like storing the objects and providing access for users similarly.

2.2.2 Thesis 2: Using the same union catalogue bears synergies referring to metadata workflows

Members of a consortium based on geographical location are likely to belong to the same library network and to use a common union catalogue. However, institutions from the same state (Bundesland) do not necessarily belong to the same network. In fact, it may even occur within the same library that diverse catalogues and metadata standards are used.

But if the partners actually do belong to the same library network, it can be useful to build on already existing infrastructure and established standards. A common standard which several partners have already agreed on usually has a twofold purpose: First, redundant work can be avoided. If a document already has an entry in the union catalogue, the other partners are able to re-use and/or extend the metadata. Second, standardization is supposed to improve interoperability and collaboration possibilities between several partners. In terms of ingesting objects into a cooperatively conducted digital preservation system by institutions which use the same union catalogue, metadata enrichment workflows are possibly the same for several partners and can be re-used, which saves personnel time and money.

2.3 Goportis' experience

As mentioned in the introduction, the Goportis consortium members all are located in Germany; but in different states.

2.3.1 Thesis 1: Being located in the same country simplifies the sharing of findings referring to legislation

The German national library and the German national subject libraries all have a collective order, the mandate, to ensure long-term-accessibility for their digital material. Hence, all three Goportis partners have to fulfill the same task, as prescribed by German legislation. Furthermore, the copyright issues are the same for every partner. Problems and tasks about copyright law can be solved in one of the three institutions, by the legal department, and the answers will be valid for all three partners. This bears synergy possibilities as the findings of the legal department of the institution are valid for all consortium members. Copyright limitations are the same for each partner, which makes it easier to formulate a common preservation policy for the Goportis consortium as well. To be subject to the same legislation has simplified consortial work for the three Goportis institutions.

2.3.2 Thesis 2: Using the same union catalogue bears synergies referring to metadata workflows

Although the three libraries are located in different states (Bundesländer) of Germany, two of them - the TIB and the ZBW - both use the union catalogue GVK of the GBV consortium, which uses PICA as a metadata standard. The ZB MED, however, uses a different one, which is based on ALEPH. For the metadata enrichment it is necessary to transform the metadata form PICA and ALEPH to Dublin Core, which is used as a metadata standard in the Goportis digital preservation system. The ZB MED developed its own metadata mapping from ALEPH to Dublin Core, which was installed separately and did not cause any problems.

The other two institutions intended to develop a common metadata mapping to use possible synergies. This, however, did not work out, as the needs and priorities of the catalogue departments of the two institutions differ too much. Lots of compromises had to be made and the different opinions on the mapping caused a time delay. In the end, the responsible staff members could not agree on some last important metadata fields and it was decided that the two institutions will have separate mappings to be able to fulfill their wishes. In the end, trying to agree on one common metadata mapping had not only not worked out but caused a time delay and much more work than it would have been if the two institutions had had two different mappings from the start. Surprisingly, this commonality has turned out to be more of a disadvantage in the end. The thesis is not supported by the experience of the Goportis consortium so far.

3. ORGANIZATION TYPE

Harold Leavitt defined an organization as "a particular pattern of structure, people, tasks and techniques" [9].In common discourse within the digital preservation context, differentiation mainly takes place at the organization purpose level, that is whether addressing businesses or cultural heritage institutions - or, to be more specific within the latter - whether talking about libraries, archives or museums. But even within an organizational purpose like "library", one needs to distinguish further. One factor is the "level", where the organization is located, whether it is an organization that operates at a national, state, city or institutional level. This is furthermore closely tied to the governance over the institution. That "level" and "governance" do not need to be congruent for comparable institutions is easily demonstrated in the case of national libraries, which can be tied to a specific governmental ministry or be independent acting branches.[1]

[1] To state a few examples: while the Library of Congress is directly administered by Congress, the National Library of New Zealand is a branch of the Department of Internal Affairs. In Germany, the national library would theoretically be part of a ministry of culture - due to federalism, such a ministry does not exist at a national level, which puts the German National

The mission of an organization is often based on "level" and "governance" - i.e. a university archive, a national museum or a state library. Size in budget, staff or collection are other ways to distinguish between organizations. A last factor, which is often overlooked but plays a big role in cooperation, is the difference in methods of operation.

3.1 Analysis of existing guidelines, best practices and reported experience

Little analysis has been done on the impact of organizational structures on digital preservation collaborations. However, it can be assumed that a large number of organizational factors influence cooperations regardless of the subject matter. No significant literature could be found describing collaborations between the industrial sector and the cultural heritage domain.[2] In a report exploring the partnerships between organizations of the cultural and educational domain, Walker et al [15] list compatibility as one of four types of risks, stating that "[...] different institutions can clash — museum curators and librarians disagree on how much and what kind of interpretive materials patrons should receive, as shown in nearly all of the digitization projects and joint exhibitions we reviewed." [15]. The other three types of risks identified are capacity, strategy and commitment. Walker and Manjarrez further describe that these risk types emerge out of 3 risk sources: innovation, complexity and institutional interdependence. The degree to which these principles are integrated into an organization will directly influence the cooperation [15]. The report suggests a number of risk mitigation strategies for collaborations (see table 1).

Table 1. Risk mitigation strategies after Walker and Manjarrez

Define clear goals and objectives	What are the projects about? What are the partners expected to accomplish?
Establish feasible timetables of tasks and deliverables	Who does what, when?
Ensure timely communication among project staff	Who knows what, when?
Make clear and appropriate project assignments	Who is responsible for what?
Recognize contribution	Who gets credit for what?
Connect like with like	Where's the right match-up across institutions?
Borrow models	Has something like this been seen before?
Accept increased risk of failure	What really counts as success when there are no benchmarks?
Create consultative mechanisms	Who should have a say, and how should they say it?
Involve senior staff in project review and decision-making	What problems require high-level resolutions?

In a call for collaborative action amongst libraries, archives and museums in the digital library domain, Liz Bishoff [4] lists the "metadata migraine" as a concrete example for risks or problems in collaborations of different organization types [4]. Different metadata standards can be seen as an epitome of problems associated with different vocabulary in cooperations. As Bishoff states, "Institutions may have common goals and visions, but they lack a common language. This lack of shared vocabulary regularly causes the professionals to talk at cross-purposes. For example, one element in a Dublin Core record is contributor. To librarians, the contributor has a role in the creation of the work - as the illustrator, translator, or photographer. To museum professionals, the contributor is a donor."[4].One benefit of inter-organizational collaborations is that of shared professional resources, offering new perspectives and insights [4].

Gibson et al explored collaborations between libraries and museums, allocating different organizational cultures and roles as the main sources of risk, which are manifested in regard to assets, personnel and professional training as well as in regards to the aforementioned terminology. Major threats derived from these risk sources are the domination of a larger partner, differences in procedure, contrasting funding sources or examples with a finer granularity, such as poor IT provision in one institution. In addition to the benefit mentioned by Bishoff [4], Gibson et al list "fostering of best practice from both institutions" and the sharing of policies [7].

It is questionable that institutions collaborate based on the factor "organizational type" alone. Usually the main drivers lie elsewhere, e.g. in similar collections, in a regional based collaboration or collaborations stimulated by a superordinate institution.

An exception to this seems to be the MetaArchive Cooperation. The foundation of the MetaArchive, which dates back to 2003, was formed by six US libraries. The organization grew into an international cooperation of different cultural heritage institutions, including libraries, museums and archives. Halbert groups most participating organizations together as cultural memory organizations, stating that "By 'cultural management organizations' I mean small to medium-sized libraries, archives, museums, and historical associations, and not enormous national agencies like the US Library of Congress or the British Library" [8]. MetaArchive forms an organizational and technological framework, utilizing a LOCKSS based infrastructure, and sees itself as "not a service provider, but a mechanism for building expertise and skills within a community-run preservation network" [16].

Communication between the members and the organization itself is facilitated through several channels: the organization itself employs a small staff-base which includes the role of the "Program Manager" and the "Collaborative services librarian". Additionally, various committees exist to address strategically and operational issues [16].

An example of an inter-organizational cooperation stimulated by a superordinate institution can be found in New Zealand. The National Library of New Zealand and the Archives New Zealand are two organizations which are comparable in governance and size, but have a different organization purpose. The organizations conduct a close cooperation in digital preservation using a joint system implementation. A joint digital preservation strategy has been written and published, describing mission and scope as well as high-level actions and role and responsibilities. One of the central purposes identified in this strategy is to "create a common understanding of digital preservation across and within the two

Library under the sovereignty of the federal commissioner for culture and media.

[2] This refers to cooperations in open-ended operative tasks and not to project-based or service based cooperations, as in the case of the development or support of a specific software

organizations" [3]. A number of "Digital Preservation Principles" were agreed upon to realize and express this common understanding. These principles include the recognition of the full preservation scope including constant management and recognition and adaption of international standards. The joint strategy leaves room for institutional decisions in regards to authenticity and integrity of the data, stating "the integrity [the authenticity] (as defined by each organization) will be retained [will be guarded and assured] through all preservation actions". Both institutions agree on not changing the original and leveraging all preservation action on a copy of the original object [3].

3.2 Theses

3.2.1 Thesis 1: Synergies through different organizational views

Organizations bring institutional knowledge and expertise into a collaboration. The knowledge and expertise can be derived from any part of the organization - its structure, people, tasks or techniques. Because no organization is like another one would assume that collaborations can benefit from the knowledge and expertise of its participants regardless of whether the organizations are of similar type or not.

3.2.2 Thesis 2: Similar organization types use similar vocabulary

Different organization types make use of different vocabulary. The literature study shows several examples where this posed a problem, for example in the form of differently used metadata fields. It should be assumed that a high similarity in organization type leads to a high similarity in vocabulary used.

3.2.3 Thesis 3: Different organizational cultures within a collaboration may form a "hidden risk"

Organizational culture demake anines how an institution works. Any organizational culture is formed by a number of factors, personnel and procedures being two of the major ones. Furthermore, organizational culture is not linked to organization type. It should be assumed that different organizational cultures within a collaboration may form a "hidden risk".

3.3 Goportis' experience

As national subject libraries, the three Goportis partners are of identical type. Furthermore, they are the only libraries of that specific type ("Zentrale Fachbibliotheken") within Germany, covering superregional, highly specialized information needs. All three partners share the same mandate of an archival library. Nevertheless, the three partners differ in many aspects, such as subjects, staffing size, holdings size or implemented technological systems.

3.3.1 Thesis 1: Synergies through different organizational views

The Goportis experience in regard to different organizational views can be broken down into three dimensions: a subject-driven synergy, an infrastructure-driven synergy and a personnel-driven synergy.

Beyond the basic scope of information providers, the procedures and furthermore the understanding of the three partners are tailored towards the needs of their respective designated communities, which in return differs based on the subject each library covers. This has a direct impact on the media and information types held in the institutions, on the way this information is presented to the respective designated community, and on overall themes of interest to the library.

As the strongest use of non-textual materials can be found in the area of science and technology, TIB places a focus on that subject matter. With the inclusion of non-textual materials - in particular AV and 3D materials - in the institutional digital preservation strategy, TIB is developing procedures which the other partners can benefit from.

An example for infrastructure-driven synergy can be found in the realized workflows. ZBW, for example, developed a submission application-passing object from ZBW's Dspace-based "EconStor" repository to the digital preservation system. The experience made there was shared with developers from the other institutions and provided valuable input for other developments.

In regard to personnel-driven synergy it has to be said that at the start of the digital preservation pilot project, the subject matter of digital preservation itself was a new task for all three libraries. Nevertheless, the three project managers - one for each library - could draw on experience from different fields of expertise (e.g. project management, information technology, research data). While this constellation of "prior experiences" may have been accidental, it proved to be very beneficial to the project. The project team managed to leverage what Walker calls "borrow models: Has something like this been seen before?" [15] in several ways: in regards to prior work experience, in regards to general procedures within their respective institutions and in regards to concrete project tasks (i.e. questions regarding tools).

3.3.2 Thesis 2: Similar organization types use similar vocabulary

As institutions of the same type with comparable procedures, especially the terminology used by the library experts needed no or little further explaining within the pilot project of the Goportis consortium. Maybe most importantly, Bishoff's "metadata migraine" [4] was not encountered. The thesis of a high similarity in organization type leading to a high similarity in vocabulary used absolutely holds true in that regard. A problem with vocabulary or terminology was, however, encountered in regards to digital preservation vocabulary itself. Concrete examples for this are terms like "preservation planning" and "risks". The partners defined procedures differently or described something as a "risk" which another partner did not see as one. This is certainly tied to the fact that digital preservation is a comparatively new task for the Goportis partners themselves, but also on global perspective - at least in comparison to well established processes like cataloguing in a library context. While the institutions are trying to connect the terminology to concrete tasks and procedures within their institutions, the terminology itself maybe in a state of slight fluctuation, so to speak. The thesis that similar organization types use similar vocabulary can thus not necessarily hold true for new practices. Based on the Goportis experience, it is advisable come to a common understanding of these terms. This can take place on a higher level, still ensuring enough room for institutional developments within a set scope of a certain term.

3.3.3 Thesis 3: Different organizational cultures within a collaboration may form a "hidden risk"

Organizational culture determines how work is conducted within the institution - hierarchy, communication style and structure are just a few examples of such influences. Not every organizational culture supports projects to allow a certain (limited) "room for

experiment".[3] Simultaneously "room for experiment" is valuable to the learning procedure for personnel and organization as a whole - especially when new tasks are concerned. Furthermore, it fosters a "cross-boundary" thinking, as it was the case in the personnel- driven synergy described in thesis one. Such outcomes are only possible if the organizations within the cooperation either employ the same "room for experiment" or are at a minimum not opposed to it within the context of the cooperation. Another example of an organizational culture-related fact is that of hierarchy and structure. Institutions position digital preservation in different positions within their organization - for some, it may be a cross-sectional task, making use of resources from different departments. For others, there is a dedicated digital preservation team or unit within a larger department. A third solution is digital preservation as a management's staff unit. The Goportis experience showed that especially for the project phase, while the institutions are still trying to define their own institutional needs for a cooperatively run system, the implementation of digital preservation as close to management as possible was extremely helpful. It formed a necessary basis for the decision on where digital preservation shall be positioned within each institution as an ongoing process. Also, the position should be clearly communicated within the cooperation, because understanding decision-making processes within partner institutions constitutes vital information.

4. COLLECTION MATERIAL

Generally collection material or subject are common reasons for libraries to collaborate:

Forming consortia to acquire similar materials is a practice often used in libraries, especially in respect of electronic resources. Advantages consist not only in a bigger market power, but also in sharing technical and legal expertise [13]. Cataloguing offers another possibility for collaboration (e.g. union catalogues, ZDB (Zeitschriftendatenbank, the world's largest specialized database for serial titles [26]).

Subject collaboration between libraries has a long tradition in Germany: the special interest collection plan of the Deutsche Forschungsgemeinschaft (German Research Foundation) supports the cooperatively distributed collection of specialized material in academic and research libraries all over Germany to meet the needs of the research community at German universities and research institutions [25].

4.1 Analysis of existing guidelines, best practices and reported experience

Expertise and technology are substantial factors of digital preservation [16]. To a large extent they depend on the library material which is to be preserved. So it seems probable that consortia with similar collection materials may benefit most from the collaboration.

But there are risks in collaborating with institutions with a similar scope of collection: Halbert [8] claims the necessity of new kinds of collaborative organizational frameworks, because (in his case) Cultural Memory Organizations are competitors for institutional prestige. In case of other institutions, competition for patrons or, directly or indirectly, monetary funds is imaginable.

[3] It has to be said that this may also of course depend on the type of project conducted.

Examples of consortia or collaborations based on common collection materials or subject are Kopal [21], PrestoCentre [24] or the North Carolina Geospatial Data Archiving Project [23].

Kopal, described above, for example developed the software koLibRI [20] to prepare the archival objects, handle the communication with the archival system used in the project and organize workflows for ingest, access and file format migration [1]. So it is evident, even if that fact isn't directly addressed in the papers, that synergies are created by working with library materials which consist of technically identical or closely related files and formats.

Shared interest in the long-term preservation of audiovisual material is the main characteristic of PrestoCentral. The PrestoCentre Foundation is a non-profit organization registered in the Netherlands under KvK54274427. It is a membership-driven organization that brings together a global community of stakeholders in audiovisual digitisation and digital preservation to share, work and learn. PrestoCentre works with experts, researchers, advocates, businesses, public services, educational organizations and professional associations to enhance the audiovisual sector's ability to provide long-term access to cultural heritage (from https://www.prestocentre.org/about-us).

The North Carolina Geospatial Data Archiving Project ran from October 2004 to February 2010. The joint project of the North Carolina State University Libraries and the North Carolina Center for Geographic Information and Analysis focused on the collection and preservation of digital geospatial data resources from state and local government agencies in North Carolina [23]. NCGDAP focused less on technical architecture than it does on partnership building and on engagement with spatial data infrastructure. The purpose of the demonstration repository II developed for NCGDAP has been: 1) to catalyze discussion within the geospatial data community about archive development, and 2) to generate learning experiences about domain-specific technical challenges associated with preserving geospatial data. To this end, a demonstration repository using DSpace was deployed, and over four terabytes of data have been acquired. A robust repository ingest workflow was developed to handle the transformation of complex multi-file, multi-formats formats into discrete digital repository items [10]. So in this project the focus was laid on the subject as well as on the formats of the preserved material.

4.2 Theses

4.2.1 Thesis 1: Similar collection materials reduce the overall costs for the collaboration

Similar collection materials enhance the positive effects of the collaboration because of synergies and sharing of technical resources and material specific experience. So in respect to the factor of collection material it makes sense to mention the cost reduction by collaborations, even if this aspect was excluded in the overall paper.

4.2.2 Thesis 2: Similar but not identical subject of collection improves the collaboration

A similar, but sufficiently different subject scope of collection addressing different groups of patrons often results in similar collection materials but avoids competition between the partners. So the collaboration can benefit from the above mentioned synergies but prevents the complications of competition for patrons or monetary funds.

4.3 Goportis' experience

As mentioned above, the three Goportis partners are of one specific type („zentrale Fachbibliotheken") within Germany, covering superregional, highly specialized information needs and sharing the same mandate of an archival library. Nevertheless the three libraries cover different subject areas with overlapping collections at peripheral areas. Their patrons benefit from the collaboration by a broader range of information and comprehensive collections even in the respective peripheral areas of collection.

4.3.1 Thesis 1: Similar collection materials reduce the overall costs for the collaboration

At the beginning of the pilot phase, the three partners concentrated deliberately on collections of the same material type (electronic dissertations and reports) in order to simplify the development of basic workflows. With growing experience and knowledge of the long-term preservation system, the institutions began to preserve technically different collections (for example Press Archives, audiovisual documents, 3D materials). Nevertheless the exchange of experience goes on and facilitates and enriches both the answers to daily problems, such as the treatment of different format types and technical interfaces, and the development of new workflows as mentioned above.

Trehub [14] describes a comparable development for the ADPNet: in the beginning all partners used identical workflows based on identical hardware to reduce costs and maintenance of the system. With the network's growing maturity the necessity of having identical hardware got less critical.

4.3.2 Thesis 2: Similar but not identical subject of collection improves the collaboration

As described above, the three Goportis partners are of the same library type with a similar scope and collections but they serve different subject needs. This fact allows the libraries to collaborate closely but nevertheless to maintain their own specific strategies for long-term preservation. As the strongest use of non-textual materials can be found in the area of science and technology, TIB by example places a focus on that subject matter developing procedures for AV and 3D materials which the other partners can benefit from.

5. Conclusion

The Goportis consortium is based on all three factors analysed in this paper; the consortium members are located in the same country, belong to the same organization type and are similar in terms of the collection material. The literature review shows that these kinds of commonalities not only make a cooperation more likely to happen, but as well simplify cooperative work and increase synergy effects and benefit possibilities for the consortium members.

As for geographical implications, commonalities like the same legislation and affiliation to the same library network with a common union catalogue enable cooperation partners to share findings and workflows more easily. The Goportis partners all belong to the same organization type, thus similar tasks and goals make synergy effects much more likely. Similarities in collection material also bear much potential for reusable workflows among each institution. Self-developed tools and plugins can be exchanged, best practice methods to solve issues with a certain kind of material can be shared and re-used.

The experience from the Goportis collaboration in digital preservation shows that the main benefits of collaboration - bundling resources to install services which might not have been possible to establish alone and reducing personnel costs by avoiding redundant work and sharing findings - have a big effect on the three German subject libraries. It would not have been possible to install an effective and efficient digital preservation system to such an extent for one of the partners alone.

Commonalities, however, can as well be false friends. Even if the pre-conditions are similar and the same infrastructure is used, the output vision can still be very different between the institutions. Driving the consortial engine on auto-pilot can easily lead to problems like failed common workflows or misunderstandings in communication. Especially in a long-term-consortium like Goportis, the daily work and the once established workflows always have to be reviewed and re-evaluated. Watching the cooperation attentively is crucial to avoid organizational blindness and to maintain a successful and beneficial consortial digital archive.

6. REFERENCES

[1] Altenhöner, R. et al. 2010. Kopal: cooperation, innovation and services: Digital preservation activities at the German National Library. In Libray Hi Tech (LHT). 28, 2 (February 2010) 235 - 244. DOI= http://dx.doi.org/10.1108/07378831011047640

[2] Altenhöner R. et al. 2011. Digitale Langzeitarchivierung in Deutschland - Projekte und Perspektiven. In ZfBB. 58, 3 (2011) S. 184-196. URL= http://zs.thulb.uni-jena.de/receive/jportal_jparticle_00237354

[3] Archives New Zealand (2011). Digital Preservation Strategy. URL= http://archives.govt.nz/sites/default/files/Digital_Preservation_Strategy.pdf

[4] Bishoff, L. 2004. The Collaboration Imperative. In Library Journal. 15, 1 (January 2004) 34-35. URL=http://www.libraryjournal.com/article/CA371048.html

[5] Caplan, P. 2010. The Florida Digital Archive and DAITSS: a working preservation repository based on format migration. In Libray Hi Tech (LHT). 28, 2 (February 2010) 224 - 234. DOI= http://dx.doi.org/10.1007/s00799-007-0009-6

[6] Diamant-Cohen, B. et al. 2003. Hand in hand: museums and libraries working together. In Public Libraries. 42, 2 (2003) 102-105.

[7] Gibson, H. et al. 2007. Links between Libraries and Museums: Investigating Museum-Library Collaboration in England and the USA. In Libri. 57, 2 (June 2007): 53-64, URL= http://www.librijournal.org/pdf/2007-2pp53-64.pdf

[8] Halbert, M. 2009. Comparison of Strategies and Policies for Building a Distributed Digital Preservation Infrastrucutre: Initial Finding of the MetaArchive Cooperative. In The International Journal of Digital Curation (IJDC). 4, 2 (October 2009) 43-59. DOI= http://dx.doi.org/10.2218/ijdc.v4i2.92

[9] Leavitt, H. J. 1962. Applied organization and readings. Changes in industry: structural, technical and human approach. In: Cooper, W.W., et al. New Perspectives in Organization Research. New York, NY: Wiley.

[10] Morris, S. et al (2010). North Carolina Spatial Data Archiving Project. Final Report. NCSU Libraries and North

Carolina Center for Geographic Information & Analysis. July 1, 2010

[11] Owen, T. et al. 1999, Libraries, museums and archives collaboration in the United Kingdom and Europe. In Art Libraries Journal. 24, 4. 10-13.

[12] The Commission on Preservation and Access and the Research Libraries Group (1996). Preserving Digital Information. Report of the Task Force on Archiving of Digital Information (May 1996)

[13] Thornton, Glenda A., 2000. Impact of Electronic Resources on Collection Development, the Roles of Librarians, and Library Consortia. Library Trends 48, 4 (Spring 2000), 812-856.

[14] Trehub, Aaron and Wilson, Thomas C, 2010. Keeping it simple: the Alabama Digital Preservation Network (ADPNet). Library High Tech 28,2, 245-258

[15] Walker, C. et al. 2004. Partnerships for Free Choice Learning: Public Libraries, Museums and Public Broadcasters Working Together. The Urban Institute and Urban Libraries Council. 22 URL= http://www.urban.org/UploadedPDF/410661_partnerships_for_free_choice_learning.pdf

[16] Walters, T. et al. 2010. Economics, sustainability, and the cooperative model in digital preservation. In Library Hi Tech. 28, 2. 259-272. DOI= http://dx.doi.org/10.1108/07378831011047668

[17] Website of the DP4Lib Project, URL= http://dp4lib.langzeitarchivierung.de/

[18] Website of the German National Library, URL= http://www.dnb.de/DE/Netzpublikationen/netzpublikationen_node.html

[19] Website of Goportis, URL= http://www.goportis.de/ueber-goportis.html

[20] Website of koLibRI, URL= http://kopal.langzeitarchivierung.de/index_koLibRI.php.en

[21] Website of Kopal, URL= http://kopal.langzeitarchivierung.de/index.php.en

[22] Website of nestor, URL= http://www.langzeitarchivierung.de/Subsites/nestor/EN/Home/home_node.html

[23] Website of the North Carolina Geospatial Data Archiving Project, URL= http://www.lib.ncsu.edu/ncgdap/

[24] Website of PrestoCentre, URL= https://www.prestocentre.org/

[25] Website of Webis, URL= http://webis.sub.uni-hamburg.de/webis/index.php/Verteilte_nationale_Forschungsbibliothek

[26] Website of ZDB, URL= http://dispatch.opac.d-nb.de/DB=1.1/LNG=EN/

ENSURE: Long term digital preservation of Health Care, Clinical Trial and Financial data

Maïté Braud
TESSELLA
Abingdon, UK
maite.braud@tessella.com

Orit Edelstein
IBM Research - Haifa
Haifa, Israel
edelstein@il.ibm.com

Jochen Rauch
Fraunhofer (IBMT)
St. Ingbert, Germany
jochen.rauch@ibmt.fraunhofer.de

Simona Rabinovici-Cohen
Kenneth Nagin
John Marberg
IBM Research - Haifa
simona@il.ibm.com

David Voets
Custodix
Sint-Martens-Latem, Belgium
david.voets@custodix.com

Isaac Sanya
Mohamed Badawy
Essam Shehab
Cranfield University, UK
i.o.sanya@cranfield.ac.uk

Frode Randers
Luleå University of Technology
Luleå, Sweden
frode.randers@ltu.se

J.A. Droppert
Philips Digital Pathology Solutions
Best, The Netherlands
aad.droppert@philips.com

Marcin Klecha
Philips Digital Pathology Solutions
Best, The Netherlands
marcin.klecha@philips.com

ABSTRACT

This paper presents the initial results of the ENSURE (**E**nabling k**N**owledge **S**ustainability, **U**sability and **R**ecovery for **E**conomic value) project, which focuses on the challenges associated with the long-term preservation of data produced by organisations in the health care, clinical trials and financial sectors. In particular the project has looked at the economic implications of long-term preservation for business, how to maintain the accessibility and confidentiality of sensitive information in a changing environment, and how to detect and respond to such environmental changes. The project has developed a prototype system, which is based around a lifecycle manager and makes use of ontologies to identify and trigger necessary transformations of the data objects in order to ensure their long-term usability. It also uses cloud technology for its flexibility, expansibility, and low start-up costs. This paper presents one of the use cases: the health care as a way to illustrate some of the challenges addressed by the ENSURE system.

1. INTRODUCTION

Ensuring long-term usability for the spiralling amounts of data produced or controlled by organisations with commercial interests is quickly becoming a major problem. Drawing on motivation from use cases in health care, finance, and clinical trials, ENSURE [1] extends significantly the state of the art in digital preservation, which to date has focused on relatively homogeneous cultural heritage data. ENSURE's use cases bring up a large number of issues, which have yet to be addressed fully, such as:

- How to leverage a scalable, pay-as-you-go infrastructure for digital preservation.
- How to get businesses to understand the economic implications of long-term preservation.
- How to create an archiving workflow that conforms to the regulatory, contractual and legal requirements of the health care, finance or clinical trials domains.
- How to maintain over the long term the integrity and authenticity of highly personal data and material covered by intellectual property rights, while ensuring access controls are respected.
- How to create a digital preservation system using only off-the-shelf IT technology.

Building on prior work, ENSURE addresses these issues with innovative approaches and tools to create a flexible, self-configuring software stack. Based on the business requirements the user enters, the solution stack will pick both the configuration and preservation lifecycle processes in order to create a financially-viable solution for the given preservation requirements, trading off the cost of preservation against the value over time of the preserved data. The main innovation areas of ENSURE are:

- *Assessment of Cost, Value, and Quality.* Ensure is creating cost, value, and quality models to help build the best preservation solution, in terms of price and performance that adheres to businesses' requirements.
- *Automation of Preservation Lifecycle Management.* Ensure uses workflow management tools to manage the execution of preservation workflows over time, thus ensuring regulatory compliance, allowing changes in the environment to be reflected in changes to the preservation approach, addressing the evolution of ontologies and managing the quality of the digital objects over time.
- *Expansion of Standard ITC Use.* Ensure is investigating using emerging technologies, such as Cloud Computing and virtualisation, to create scalable and financially-viable solutions for long- term digital preservation.

- *Creation of Content-Aware Long-Term Data Protection.* Ensure is researching how to secure data over the long-term, when that data is affected by new and evolving regulations, contains personally-identifiable information, and needs to be accessed by a changing user community with differing roles.

The ENSURE project started in February 2011 and has created a reference architecture already and demonstrated many innovations in its initial implementation.

Section 2 presents the overall architecture of the ENSURE system, section 3 and 4 describes the two main components of the ENSURE system: the Configuration Layer and the Runtime System, Section 5 present a use case, and section 6 gives our conclusions..

2. ENSURE ARCHITECTURE

The ENSURE system's architecture consists of:

- A set of plug-ins that provide specific functionality such as format management, regulatory compliance, integrity checks, and access to specific storage clouds.
- A runtime Service-Oriented Architecture (SOA) framework that allows an OAIS [2] solution to be created from those plug-ins needed to meet a user's requirements, including any economic considerations (s)he has.
- A configurator and an optimiser which use cost/quality analysis engines to create and evaluate a proposed preservation solution.

A high-level view of the ENSURE architecture is given in Figure 1, which shows that there are two layers: the *Configuration Layer* and the *System Runtime*.

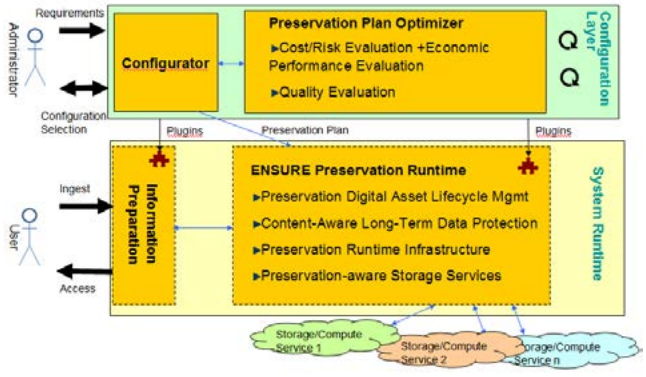


Figure 1. ENSURE System Overall Architecture

These two layers are described in the next two sections.

3. CONFIGURATION LAYER

The components in the *ENSURE Configuration Layer* are run prior to the initial deployment of the preservation solution and are re-run periodically; in particular they need to be re-run if there are major environmental changes. These components create the preservation plan used by the preservation solution in the first place and update it when the environmental or business needs change.

The architecture of the Configuration Layer is given in Figure 2.

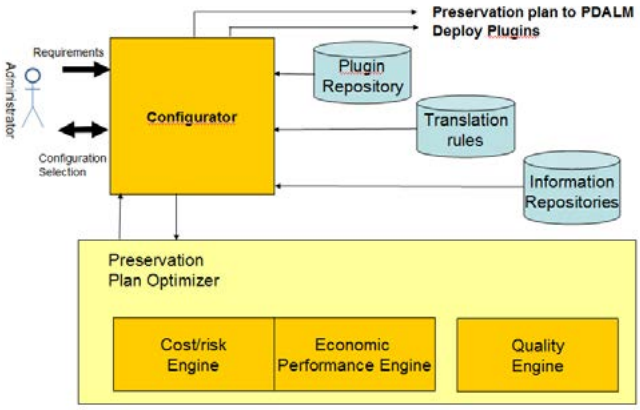


Figure 2. Configuration Layer Architecture

The flow of operation in the Configuration layer is as follows:

- The administrator is presented with a form and enters the business' requirements and preferences for the preservation system. When the system is being reconfigured, the previous configuration is shown to the administrator for reference.
- Using a Rule Engine and a set of rules, the Configurator constructs a parameterised global preservation plan (GPP) consisting of a preservation plan and its associated configuration. A global preservation plan (GPP) defines where and how data will be preserved; this includes encryption of data, fixity checks and storage provider. A parameterized GPP describes a collection of potential plans by means of parameters that take values from well-defined ranges. For example, one parameter could define a collection of possible encryption algorithms, and another parameter could define a collection of storage providers.
- The Preservation Plan Optimiser (PPO) explores the collection of potential plans, returning to the Configurator a small number of plans that are optimised with respect to cost and quality. The PPO uses the Quality Engine and the Cost Engine to provide evaluations of potential plans. These evaluations drive the optimization.
- The Configurator presents the top three preservation plans to the administrator together with their evaluations. Either the administrator can select the solution to deploy, or (s)he can request a modified configuration, which will restart the process.
- When a preservation solution is chosen, the Configurator deploys it by:
  1. Deploying the selected plug-ins in the runtime infrastructure and activating the associated services in the appropriate environment.
  2. Activating the Preservation Digital Assets Lifecycle Management component and passing it the preservation plan.
  3. Storing the selected configuration and its evaluation in the ENSURE system in order to preserve it.

3.1 Preservation Plan Optimiser

Finding preservation plans that are optimised with respect to cost and quality is a multi-objective optimisation problem. Typically the objectives are conflicting and there is not a single best solution. Evolutionary algorithms are widely used to find solutions that are Pareto optimal [3]. The PPO uses the evolutionary algorithm NSGA-II [4] to explore the collection of potential plans and find optimal solutions. It defines a genotype that encodes the parameters of the parameterised GPP. For example, there can be a gene representing a choice for an encryption algorithm. The evolutionary algorithm selects actual values for the genes of the genotype, thus generating candidate plans that PPO then sends to the engines for evaluation of quality and cost. The quality and cost values thus obtained act as objective values that are maximised or minimised in the optimisation performed by the evolutionary algorithm.

Several software frameworks exist that provide implementations of evolutionary algorithms. The Opt4J optimisation framework [6] has been selected for the PPO.

In order to take account of user preferences, the ENSURE project uses *a priori* preference articulation i.e. the user expresses preferences before the optimisation is performed. The PPO defines a weight function on the objective space to represent the user's stated rating of the importance of the different objectives. The Opt4J implementation of NSGA-II has been extended to use such weightings, as described in [5]. The selection performed by the evolutionary algorithm thus favours solutions that score well on the objectives that the user considers important.

3.2 Cost Modelling for Long-Term Digital Preservation

Assessing the cost and economic value of preserving digital information is important for organisations performing preservation activities. Therefore, one of the aims of ENSURE is to develop a cost model and a cost engine to predict the 'whole life-cycle cost' of LTDP in the cloud. The developed cost model will focus mainly on three business sectors: healthcare, financial and clinical trials.

The core activities involved in the design and development of the cost engine for ENSURE include:

1. Identification of the work break down structure (WBS) and cost break down structure (CBS) of digital preservation activities, as identified in Figure 3.
2. Identification of cost drivers, risks/uncertainties factors, and obsolescence issues in LTDP activities
3. Development of the cost model, including implementable cost equations and rules.
4. Implementation of the cost engine as a web service and its integration into the rest of the ENSURE architecture.

The activity based costing (ABC) methodology has been employed to develop the cost model. The ABC approach enabled the development of a generic cost model that is applicable to and relevant for not only the ENSURE use-case organisations but also other industries. The cost model is translated into a set of equations and rules to enable the accurate estimation of cost for LTDP activities.

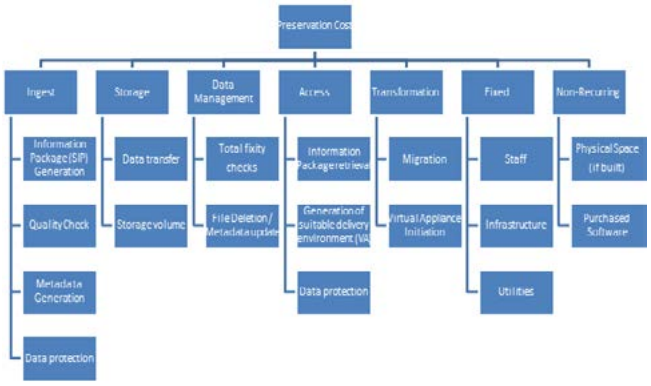


Figure 3. ENSURE Cost Breakdown Structure of Digital Preservation Activities

3.2.1 Challenges in Cost Estimation for LTDP

For ENSURE, the main challenges of estimating the cost of LTDP are as follows:

- Long-term digital preservation is being applied to three new business sectors. Most previous work in digital preservation has focused on the science and cultural heritage sectors.
- There is no established definition of uncertainty for LTDP.
- No research on the impact of uncertainty on cost has been undertaken and information about this topic is scarce.
- Limited work has been done to investigate the cost of ameliorating obsolescence through LTDP.
- LDTP made use of cloud computing only recently, so cost data is scare. Cloud costs are split between cloud storage and cloud computing.
- Determining the cheapest configuration is made harder by the number of parameters that can be optimised.

3.2.2 Cost Engine Architecture

The cost engine system architecture comprises several communicating components that implement the overall ENSURE cost evaluation and optimisation system. Figure 4 illustrates the architecture of the cost engine and how its modules interact with the rest of the ENSURE system. The GPP describes aggregation-specific (see section 4.5.2) (e.g. encryption, fixity, etc.) and copy-specific (e.g. storage, computing) preservation actions and the preservation configuration. The preservation configuration describes the physical architecture, software, and plug-ins employed for digital preservation activities. The cost engine results include initial investment cost, year one cost, ingest cost, data management cost, storage cost, access cost and reconfiguration cost for the data retention period (given in years) in the configurator.

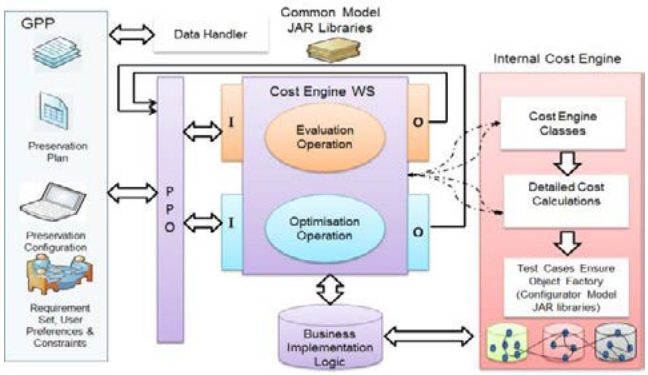


Figure 4. ENSURE Cost Engine Architecture

3.2.3 Validation

The cost engine has been validated qualitatively via expert opinion in the digital preservation community. The first phase of validation was the cost break down structure, followed by the equations and rules that have been implemented. There are plans to validate the cost engine quantitatively with real cost values to ensure its generalisability, applicability and validity.

4. RUNTIME SYSTEM

The ENSURE System Runtime is the SOA infrastructure for executing the plug-ins selected by the Configuration layer. This layer provides data management and archival storage services, as well as ingest and access services. It interacts with external storage services which provide the physical space for storing the preserved data and potentially it interacts with the external compute service, which runs in the storage layer to minimise i/o overheads. In addition, this layer watches for environmental changes that may require the system to be reconfigured.

The architecture of the runtime system is given in Figure 5.

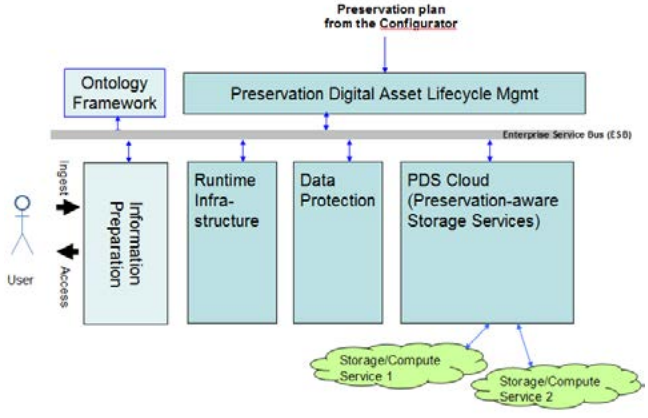


Figure 5. Runtime System Architecture

The components of the Runtime System are:

- Preservation Digital Asset Lifecycle Management: It manages the workflow of the information being preserved by executing the preservation plan built by the Configurator. In addition, it manages the system's log and all provenance information. Furthermore, it handles changing the workflow at reconfiguration, and it monitors internal and external events including watching the environment for events that need administrator attention or reconfiguration. Also, it handles sending out notifications and all interactions with the administrator.
- Information Preparation: It runs when the data is being ingested or accessed. Upon ingest, it prepares the information and metadata ready to be preserved, generates the search indexes, and packages the data. Upon access, it handles locating the data in the index and packaging it for the user. Its data protection functions are used to ensure access rights are observed.
- Ontology Framework: It manages the preservation Ontologies and search Index. It also supports the evolution of the Ontologies.
- Preservation Runtime Infrastructure: It evaluates the quality of the managed information, supports data transformations, and supports a range of approaches for future accessibility.
- Preservation-aware Storage Service: It stores the digital resources in external storage services using cloud storage, validates the bit-level integrity of the data, manages provenance at the storage level, and supports running computations in the cloud storage layer.
- Content-aware Long-Term Data Protection: It ensures that the use of sensitive information over the preservation life-cycle complies with the specified long-term access controls, privacy restrictions, IPR protection rules, and de-identification, and anonymisation requirements.

4.1 Preservation Digital Asset Lifecycle Management (PDALM)

ENSURE has researched the integration of existing approaches to Lifecycle Management with digital preservation and this research is encapsulated in the PDALM component. It orchestrates the management of an asset from ingest to disposal, by invoking components developed in other work packages. Objects are disposed of only according to the applicable rules and regulations of the relevant business sector, together with the relevant business objectives.

In essence the PDALM component is the "brain" of the ENSURE system and therefore is also responsible for controlling the system activities, and handling notifications to and interactions with the administrator.

4.1.1 Workflow Engine

The PDALM component is principally a workflow engine, which is capable of running those workflows whose details are specified in the Preservation Plan created by the Configurator. It is capable of starting workflows manually or automatically (based on timers or pre-defined rules). The workflow types are consistent with the OAIS model: Ingest, Access, Preservation, and Data Management workflows are available.

The workflow steps themselves are not executed within the workflow engine, but it is responsible for sending web service requests to the other runtime components (Information Preparation, Data Protection, PDS Cloud) to execute the workflow steps.

The workflow engine is based on the open-source workflow engine jBPM (released by the JBoss Community under the ASL license). jBPM comes with a web-based console that allows the user to start workflows and control running workflows. This jBPM console forms the basis for the PDALM Graphical User Interface (GUI).

The PDALM workflow engine contains a component responsible for translating the Preservation Plan created by the Configurator into a series of Ingest, Access, Preservation and Data Management workflow definitions, which will be uploaded automatically into the workflow engine.

A key point to reacting to changes is the ability to reconfigure the system. In collaboration with the Configurator component, the ENSURE system is capable of reconfiguring a running instance of the workflow engine. There are two types of workflows available in the workflow engine: manual workflows which are started by the user, and scheduled workflows which are started automatically by the system. In addition, when the system is up and running there may be workflows waiting for a manual or timed trigger (e.g. transformation workflow) as well as running workflows. In order to reconfigure the live system, the PDALM component needs to be able to put the workflow engine in a dormant state, which is done in stages. When the Configurator notifies the PDALM component that a new preservation plan needs to be deployed, the PDALM component stops all the workflows that are either waiting or scheduled to run, before waiting for all the active, running workflows to complete; only then is the system in a dormant state. Once in this state, the workflow engine can be stopped safely and reconfigured based on the new preservation plan. If still required, the workflows that were scheduled to run or were in a waiting state can be restarted in the newly reconfigured system.

4.1.2 Event Engine

The PDALM event engine is responsible for monitoring internal and external events relating to environmental changes that could affect the long-term preservation of the data preserved by the ENSURE system. Monitoring external events is a difficult task as the source of such events can be as diverse as the ways of monitoring them.

Initially effort was focussed on one of the pre-existing central repositories of information for long-term preservation: PRONOM. PRONOM is developed and maintained by the UK National Archives (TNA) and holds impartial and definitive information about the file formats, software products and other technical components required to support long-term access to electronic records and other digital objects of cultural, historical or business value.

One of the current limitations of PRONOM is that as its information is stored in a relational database, it is difficult to update or merge two instances of PRONOM. Also it makes it difficult to identify which information has changed when an instance of PRONOM is updated. To solve these problems, in 2011 TNA started to implement Linked Data Pronom with the plan to release the data held by PRONOM in a linked open data format in order to make it easier to reuse. Such a Linked Data registry makes it easier to compare two instances of the same registry and detect if and what changes have occurred.

The ENSURE project has started to extend this by adding additional Linked Data instances to complement the information held by Linked Data PRONOM including information in PRONOM but not yet by Linked Data Pronom and also information relevant to the ENSURE use cases, such as cost, hardware, and data protection. The resulting Linked Data network will consist of, for example, external data held in the Linked Data PRONOM instance maintained by TNA, data held in a Linked Data instance about other relevant technical information (e.g. local tool capabilities), and data held in a Linked Data instance about costs maintained by ENSURE. This Linked Data network will help to demonstrate how it is possible to get notification of external events and react to them by using Linked Data.

Apache Jena was chosen as the Framework to handle the Linked Data network as it provides the following functionality out of the box:

- an API for reading, processing and writing RDF data in XML, N-triples and Turtle formats,
- an ontology API for handling OWL and RDFS ontologies,
- a rule-based inference engine for reasoning with RDF and OWL data sources,
- stores to allow large numbers of RDF triples to be stored efficiently on disk,
- a query engine compliant with the latest SPARQL specification, and
- servers to allow RDF data to be published to other applications using a variety of protocols, including SPARQL.

The event engine contains functionality to query the Linked Data Pronom instance maintained by TNA; a scheduled BPMN workflow, running within the PDALM workflow engine, queries the distant Linked Data Pronom instance and compares it to a snapshot of stored locally in order to detect and identify any changes that might have occurred in since the last query. Then the impact of the change is calculated using the Linked Data network presented above and communicated to the administrator of the ENSURE system via email. Given this information, the administrator may choose to request that the Configurator calculates a new Preservation Plan.

This is illustrated in the following example: The date that the creator of a file format will withdraw support is updated in TNA's Linked Data PRONOM instance (or a copy of this instance) and this change is detected when the scheduled comparison workflow runs. This change is detected and triggers looking up a preservation action to perform as a consequence (e.g. format migration). Then, the event engine will calculate the financial impact of the change using the data stored in a further part of the Linked Data Network. In this case, therefore, the cost will be the cost of running the tool and the additional cost of storing the migrated data based off the chosen migration strategy triggered from the detection of the external event of obsolescence of the file format.

Much work in many different initiatives is being undertaken to unify technical registries and other repositories of digital preservation information: e.g. UDFR[24] and LDS^3[23] are focusing on using semantic web and Linked Data to enable the sharing of information. Therefore the Linked Data registry developed as part of the ENSURE project will not be limited to

linking to Linked Data Pronom only but will be capable of linking to other Linked Data registries as well, provided that their vocabulary specification is published and freely available.

4.2 Preservation Information Preparation

Information preparation plays an important role in any digital preservation system as it has to ensure that during ingest all the necessary information required for preservation, long-term accessibility and usability of the data objects to be preserved is gathered. The OAIS reference model reflects this both in the Ingest and Access components and in the different information packages of the OAIS information model that are produced or processed by the Ingest and Access component, namely the Submission Information Package (SIP), the Archival Information Package (AIP) and the Dissemination Information Package (DIP). The ENSURE project has demonstrated that it can ingest and retrieve simple and more complex data objects (e.g. DICOM image file sets) from its three target domains. Figure 6 illustrates the data workflow in the Information Preparation architecture. Ingest services were developed that select the right information to be preserved from the test data for each use case, extract the metadata relevant to the data objects' MIME types , package everything in an AIP and hand it over to the ENSURE Preservation Runtime for preservation. Furthermore, access services were developed to provide efficient search and retrieval of data objects in the form of DIPs. In particular, semantic web technologies were applied to model, collect and manage the metadata of the digital objects from the different domains effectively and to provide a powerful search and access mechanism for preserved data. By representing the Data Objects' metadata in terms of an integrated set of formal ontologies, the preservation knowledge and domain-specific object formats and concepts can be modelled in an application-oriented way. The ontologies contain concepts describing the general features of Data Objects (i.e., type, format, size, Preservation Description Information) as well as domain-specific information. The captured metadata of the Data Objects represent instances of the ontologies and are encoded as RDF triples and stored by the ENSURE Preservation Runtime in an index.

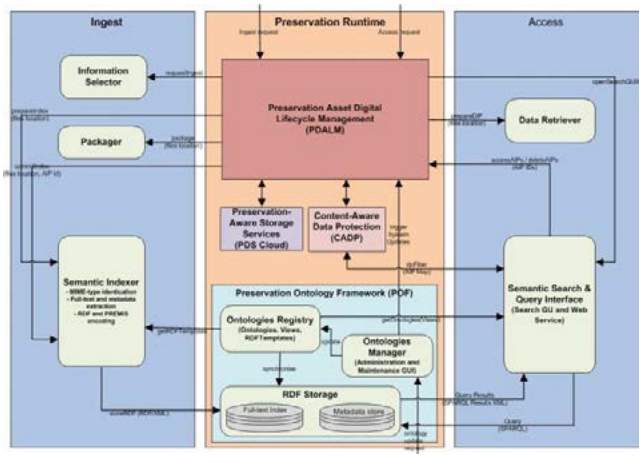


Figure 6. Overview of the Information Preparation Architecture

4.3 Ontologies Framework

The integrated Preservation Ontology Framework (POF) includes an Ontologies Registry and a set of ontologies related to preservation, which is provided as a subset of the Nepomuk Information Element Ontology (NIE) [7]. This provides the flexibility required to serve the unknown, future data retrieval needs of the user community. It provides the platform to research how the evolution of the ontologies over time can be managed in an archive, and can be exploited to identify and trigger necessary transformations of the data objects in order to ensure their long-term usability. Further, it enables investigations into how the knowledge coded in ontologies can be used to resolve other preservation-related problems, such as the protection of sensitive healthcare data under changing regulations. To do so, a management component, the Ontologies Manager, was implemented to enable the user to maintain different versions of ontologies through a GUI (see Figure 7). In addition to managing the update of ontologies, the Ontologies Manager executes any system adaptations necessitated by the creation of a new version of an ontology, such as re-indexing archived AIPs in order to keep the entire system consistent. The COnto-Diff algorithm [8] is used to calculate the differences between sequential versions of ontologies and provides both the information required to execute the necessary system adaptations and an estimate of the required effort.

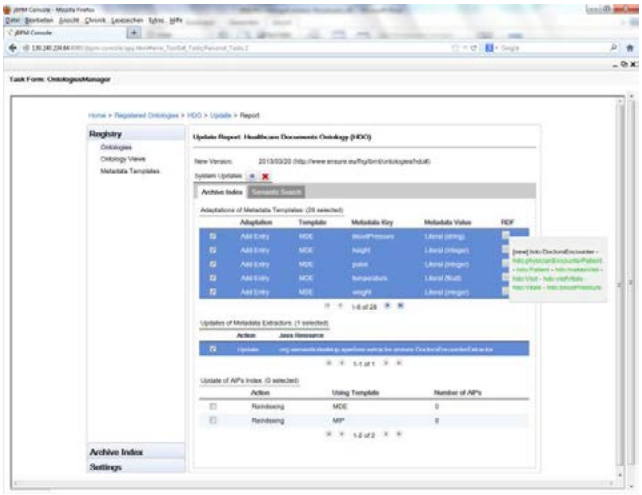


Figure 7. Screenshot of the Ontologies Manager

4.4 Preservation Runtime Infrastructure

The Preservation Runtime Infrastructure which is part of the Preservation Runtime supports a range of approaches to future accessibility including both transformation and emulation/virtualisation. This component is responsible for providing the transformations of formats, for evaluating the usability of the information after a transformation, and for periodically evaluating the quality of the managed information. This section describes how the quality of the information can be assessed.

In part a system for long-term digital preservation of information can be viewed as a communication system, sending information from a producer to a consumer through a channel (the preservation system). Unlike the channels encountered in standard communication systems, this channel has an extreme intrinsic time delay – possibly measured in decades – that makes any type of feedback-loop impracticable, if not impossible. Under very specific circumstances it is possible to use information theory to analyse the effect of a specific use of an information

system [12] (such as a communication system), but since it is not possible to represent mathematically the types of uncertainties that encountered in digital preservation, it is not possible to use information theory to study the effect of digital preservation systems in generalised use [13]. As the information transfer in digital preservation is determined not only by input and output symbol alphabets and their conditional probabilities, but also depends to a great extent on pre-knowledge and qualitative factors, the authors are forced to conclude that it is not possible to model the digital preservation ”channel” using traditional information theory [14].

What makes the digital preservation domain so elusive and hard to capture in strictly technical terms is the extent to which qualitative factors, such as trust and authenticity, influence the perceived quality of the transferred information. It could be argued that the rendering an image in an obsolete format using emulated viewing software does not differ from migrating that image and then viewing it using contemporary software, but they do differ in terms of the amount of trust you need to have: trust in the chain of migration software used to keep the image up to date and trust in the organisation managing the process [9][10] versus trust in the emulator and the process used to select it.

The ENSURE system aims to empower the preservation services customer (the producer in OAIS terms [2]) to choose an appropriate preservation plan for two reasons: (1) the choice affects the cost and thus should be taken by the customer, and (2) the customer is best equipped to assess the qualitative impact of the proposed preservation plans.

Digital preservation is not only a set of technical problems related to technology, formats and algorithms, but also a problem that concerns the interface between technological systems and humans. Most of all it is a problem concerning the mutual understanding between humans separated in time – and thus by culture. In the ENSURE setting, the preservation organisation aims to help the producer to understand the effects of the chosen preservation plan on the predicted needs of the future consumer and how to best fulfil these needs within the available budget.

There is a fundamental conflict in demanding a decision from the producer regarding a proposed preservation plan because at least two conflicting concerns govern the actions of the producer; minimising the cost and maximising the quality of the transferred information. It is inevitable that the producer and the consumer (and in the ENSURE case the consumer is the producer at a later point in time) will have different views of the information and the use of that information [11] (see Figure 8).

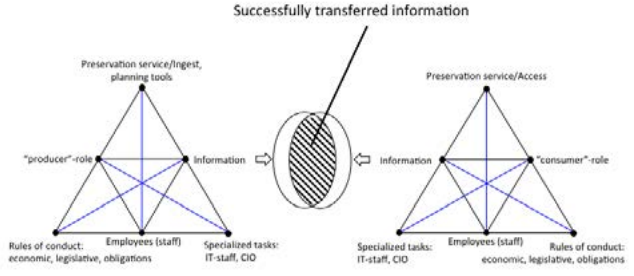


Figure 8. The Producer and Consumer Perceive the Information Differently, c.f. [4]

In order to help the producer make an informed decision, ENSURE equips the producer with a tool that supplements the cost of choosing a specific preservation plan with the consequences of choosing that plan. The consequences are provided as (1) the monetised cost of risks based on calculations of economic performance, (2) a metric assessment of perceived quality from a predicted consumer viewpoint, and (3) a set of qualitative statements of the failure to exhibit specific characteristics of quality [17]. These consequences are predicted by attempting to extrapolate the current usage of the information into the future based on the assumed purpose of use of the information together with the purpose of preserving the information.

Empirical data gathered through a series of interviews with the use case owners in ENSURE emphasises the differing concerns of businesses needing preservation and the organisations providing preservation services. Businesses often lack knowledge of digital preservation, while preservation service providers often struggle to understand the specific needs of those businesses requiring their services. As these two organisations are effectively working together to predict the future needs of the business organisation, it is essential that they communicate effectively. This communication has to be based on a shared mental model that is expressive enough to capture the immediate needs of both organisations [16][15].

4.5 Preservation-Aware Storage Service

Preservation Data Stores in the Cloud (PDS Cloud) is an OAIS-based [2], preservation-aware, storage service in a multi-cloud environment. Unlike existing cloud storage systems, or even some traditional archival systems, PDS Cloud supports logical preservation; in addition, it converts logical preservation information objects into physical cloud storage objects. The idea behind PDS Cloud is that digital preservation systems will be more robust and will reduce the probability of data corruption or loss if preservation-related functionality is offloaded to the storage system.

The foundations of PDS Cloud were established in PDS [18], a preservation storage architecture using Object Storage Devices (OSD). For the ENSURE system the scope has been expanded and adapted for the cloud environment. The following cloud-specific goals and requirements have been added:

- Support access to multiple cloud storage and cloud computing platforms, and enable migration of data between different clouds. This includes using multiple clouds concurrently, while taking advantage of the special capabilities of each platform.
- Provide a flexible data model for a multi-tenant, multi-cloud environment, with easily configurable data management capabilities that can be tailored for diverse aggregations of digital assets having different preservation requirements that can change over time. A key feature is the ability to change the physical placement of objects in the cloud without affecting how the user accesses the data.
- Enhance the future understandability of content by supporting data access using cloud-based virtual appliances. Each virtual machine instance is created from a previously published image or from readily available components and provided with the desired preservation data content and the designated software needed to render the data.

- Offer advanced OAIS-based services, such as fixity (aka integrity) checks, provenance records and auditing services that complement the generic cloud's capabilities. Also, it must support complex, interrelated objects and manage their relationships and links while maintaining referential integrity.

While this section provides an overview of the architecture and data model of PDS Cloud, a more comprehensive presentation of the PDS Cloud system can be found in [19].

4.5.1 Architecture

PDS Cloud is designed as an intermediate service layer, providing a broker that connects the OAIS entities with the multiple cloud systems; in addition it fulfils the role of the Archival Storage component in the OAIS functional model. PDS Cloud exposes a set of OAIS-based services, such as ingest, access, deletion and preservation actions [2], to the client and uses heterogeneous storage and computing cloud platforms from different vendors. AIPs may be replicated to multiple cloud storage systems to exploit different cloud storage capabilities and pricing structures, and to increase data survival.

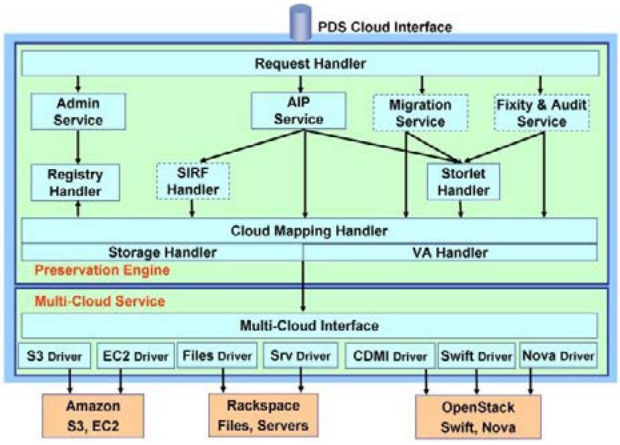


Figure 9. PDS Cloud High-level Architecture

As shown in Figure 9, PDS Cloud is divided into two main layers: a Multi-Cloud Service, and a Preservation Engine.

- *Multi-Cloud Service:* This handles access to a heterogeneous set of cloud storage and computation platforms. Its role is to encapsulate the specific interfaces and capabilities exposed by each different cloud platform. It is agnostic to preservation and is implemented using jclouds [20], an open source cloud interface library that comprises a unified interface (multi-cloud interface component) and a set of drivers that implement the interactions with the individual storage and computation clouds underneath.
- *Preservation Engine:* This provides the preservation functionality for AIPs. It receives requests from PDS Cloud clients and services them using various functional handlers organised in several levels. At the top level is the Request Handler, which is the server side of the HTTP interface. When it receives an HTTP request, it validates it, before handing it over to the appropriate handler for processing. At the lowest level is the Cloud Mapping Handler, which maps AIPs to the cloud object model, and interacts with the Multi-Cloud Service layer to perform operations in the cloud.

This architecture, with its separation of concerns, is designed to support the deployment of multiple clouds from different vendors. Providing such heterogeneity allows the user to experiment with diverse technologies and to determine whether appropriate actions have been taken to ensure continued access to the AIPs despite the diversity of current technologies; this is analogous to ensuring continued access to AIPs despite the change in technologies over time, i.e. ensuring their preservation.

4.5.2 Data Model

Users should be able to access their data without needing to know the details of how or where it is stored. PDS cloud hides the complexity of a dynamically configured, multi-cloud, multi-tenant environment behind a simple facade that uses a uniform, hierarchical resource naming path for entities and an abstract data model that allows for multiple implementations to take advantage of the different capabilities of the cloud storage platforms being used.

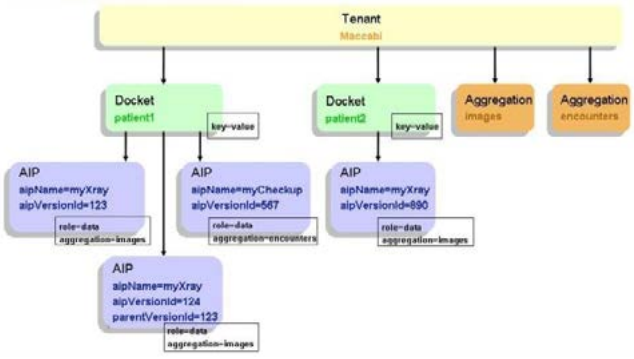


Figure 10. PDS Cloud Data Model

The data model, which is illustrated in Figure 10, comprises four types of entities: tenant, aggregation, docket and AIP.

- *Tenant:* This entity is an enterprise or organisation that engages in storing data in the cloud. Each tenant constitutes an independent information domain, which has separate administrative ownership, policies and users. Data assets belonging to different tenants are isolated logically from each other.
- *Aggregation:* This entity is a configuration profile, which defines the policies and capabilities for managing the data in storage. It specifies the details of one or more cloud platforms (address, credentials, etc.) that are being used for physical storage. It also designates various characteristics for maintaining and accessing data, such as integrity checking procedures or rendering properties that are relevant for the specific use case.
- *Docket:* This entity is a grouping of preserved entities; it is analogous to a directory in a file system.
- *AIP:* This entity is the fundamental preservation entity in OAIS. An AIP has a name (aipName), as specified in the hierarchical path, a logical identifier (aipLogicalId), and a version identifier (aipVersionId). Multiple versions of the same AIP are distinguished by their aipVersionId, while the

aipLogicalId is common to all versions of the same AIP. The combination of aipLogicalId and aipVersionId is globally unique. When an AIP is moved to a different docket, its Ids remain the same, this enabling continued access. Each AIP is associated with a specific aggregation, and is replicated to all the storage clouds configured in the aggregation.

Aggregations are configured based on the needs of the tenant. The AIPs in a given aggregation can be viewed as a collection of information assets that share the same characteristics and are managed together and in the same fashion. In that sense, aggregations can be considered as part of the service layer.

Users can access their data without needing to know the configuration details held in the aggregation. It is the responsibility of the storage service layer, i.e. PDS Cloud, to interpret the aggregation's configuration profile in order to access the specific cloud platform(s) and map the logical dockets and objects to the physical name space of each specific cloud. So although changes to an aggregation's configuration over time affect how it is handled by the storage service layer, they do not affect the user application interface.

4.6 Content-Aware Long-Term Data Protection

The preservation of data over long timeframes poses a series of unique problems for the protection of sensitive content. This includes both commercially valuable data and that covered by current and future data protection legislation. Initially ENSURE developed a set of scenarios focussing on the effects on data protection of changes in the legal, social, political, and technological landscape over long periods of time and then considered how to address these threats. For example, in order to deal with technological issues such as the cracking of an encryption algorithm, the system is designed with an automatic re-encryption method to preserve confidentiality.

A key challenge is in the design of a future-proof access control mechanism. The mechanism must be flexible enough to support different types of access control models (such as role-based access control (RBAC) or lattice-based access control (LBAC)) and must be able to deal with syntactic and semantic changes (e.g. changes in file formats, application domain concepts, user roles, etc.). Given these constraints, an access control engine that implements the OASIS XACML v2 specification, including hierarchical resource profiles [21] and multiple resource profiles [22] was chosen. By supporting these profiles, access control policies for hierarchical resources can be combined into a single authorisation decision, thus simplifying access control policy management, and content filtering can be applied to DIPs so that only those parts of the original AIP to which the requesting user has access are delivered to him/her. The engine was extended to support hierarchical subject attributes (i.e. hierarchical roles) and to support concepts of purpose of access. The latter are not supported by the standard XACML specification but have been identified as necessary to support the access control policies identified in the use cases.

In order to make writing access control and privacy policies as simple as possible, the RDF Storage component and ontology framework described in Section 4.3 is relied upon extensively. Which policies apply for a given authorisation request can be determined by exploiting the metadata produced by the Information Preparation component. Furthermore, policy rules can refer to attributes that are not specified as part of the access control request (e.g. attributes related to resource content, such as which patient a medical record pertains to). Policies may also rely on the ontology framework for classification of security and privacy-related concepts and to deal with potential changes in domain-specific concepts that might otherwise require access control policies to be rewritten.

A plug-in based obligation handler framework, which is integrated into the authorisation engine, is relied upon to deal with encryption, de-identification and other security and privacy obligations specific to the application domain.

The governing access control policies are ingested in the same way as regular AIPs and can be interchanged or updated at run-time, thereby changing the policies that are in effect and making the data protection system future-proof.

5. HEALTH CARE USE CASE EXAMPLE: DIGITAL PATHOLOGY

Long-term data preservation is vital for the Healthcare domain, just as it is for the sub-domain of Digital Pathology. The term Digital Pathology is used to describe the current trend amongst pathology departments to digitise pathology glass slides.

As the pathology glass slides are scanned at high resolution, a large amount of data is generated, up to 2TB per day, and this data must be preserved. As well as storing the digitised glass slides (aka Whole Slide Images (WSI)), the preservation system must store other objects, such as documents (e.g. reports, order forms), the results of image analysis applications, as well as the corresponding case, patient, and slide metadata. These different objects will be stored in the Digital Images and Communications in Medicine (DICOM) format.

As multiple sites may work and store data for the same customer, it must be possible to access the central preservation system from all these sites. Typically, newly created images and objects are likely to be accessed more frequently (e.g. for QA review) than older ones. Pathologists may retrieve images and objects via a digital pathology application, or patient histories, which could be relevant to the diagnosis of a recent case, directly via the preservation system's web interface. While researchers interested in developing image analysis applications or carrying out data analyses for scientific purposes also need access.

All users require real-time search and browsing of the preservation system, but not all users have the same access rights to all the objects. Therefore the system supports user authentication and authorisation with role- or profile- based access to the objects, in order to protect patient privacy.

As part of the image life-cycle management, it should be possible to transcode, transform or process images without exporting the entire (large) image files from the preservation system, in order to, reduce the cost of storage or save bandwidth. The ENSURE system can store image analysis applications as Virtual Applications and therefore offer image processing close to the stored data. Transformations, whether implemented as automatic or manual workflow steps, may be triggered by external events that the preservation system is configured to watch for, such as changes in regulations that require either a longer or shorter mandatory image storage period. Therefore, the preservation

system needs to allow images and other objects to be updated in a controlled manner.

6. CONCLUSION

In this paper we have provided a general description of the ENSURE system which aims at helping organisations in the health care, clinical trials and financial sectors to prepare and evaluate cost effective preservation plans and build corresponding flexible archival systems based on a set of available plug-ins. The system stores the organisation's content in public or private clouds while maintaining protected access to sensitive data. In addition, it supports a set of preservation-related ontologies, which provide a flexible and future-proof way to search for archived data. Finally, its technical registry, which is based on Linked Data and connected to external registries, is capable of detecting environmental changes that might affect the archival system and require a change in the preservation plan. Such reconfiguration of a live system is also supported. A prototype system implementing the presented architecture has been developed during the two first years of the project and its evaluation by partners in the health care, clinical trials and financial sectors is planned for the last year of the project.

7. ACKNOWLEDGMENTS

The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement n° 270000 - ENSURE.

8. REFERENCES

[1] ENSURE, DOI=http://ensure-fp7.eu/

[2] The Consultative Committee for Space Data Systems (CCSDA). 2012. Reference Model for an Open Archival Information System (OAIS) – Recommended Practice, CCSDS 650.0-M-2 (Magenta Book) Issue 2. Also available as ISO Standard 14721:2012. http://public.ccsds.org/publications/archive/650x0m2.pdf.

[3] Coello Coello, C.A., Lamont G.B., Van Veldhuizen, D.A. 2007. Evolutionary Algorithms for Solving Multi-Objective Problems. Springer, New York, USA, second edition, September 2007. ISBN 978-0-387-33254-3.

[4] Deb, K., Pratap, A., Agarwal, S., Meyarivan, T. 2002. A Fast and Elitist Multiobjective Genetic Algorithm: NSGA-II. In *IEEE Transactions on Evolutionary Computation*, Volume 6 Issue 2, April 2002. IEEE Press, Piscataway, NJ, USA.

[5] Friedrich, T., Kroeger, T., and Neumann, F. 2011. Weighted Preferences in Evolutionary Multi-Objective Optimization. In *AI'11: Proceedings of the 24th international conference on Advances in Artificial Intelligence.* Springer. LNCS 7106.

[6] Lukasiewycz, M., Glaß, M., Reimann F., and Teich J. 2011. Opt4J – A Modular Framework for Meta-heuristic Optimization. In *GECCO'11: Proceedings of the 13th annual conference on Genetic and Evolutionary Computing* (Dublin, Ireland, July 2011). ACM, New York, NY, USA.

[7] NEPOMUK – Social Semantic Desktop, funded under FP6-ICT Programme, DOI= http://nepomuk.semanticdesktop.org/nepomuk/.

[8] Hartung, M. and Gross, A. and Rahm, E.. 2013. *COnto-*Diff: Generation of Complex Evolution Mappings for Life Science Ontologies. *Journal of Biomedical Informatics.*

[9] Duranti, L. 1989. *Diplomatics: New uses for an old science, part I Archivaria*, 1(28).

[10] Duranti, L. 2002. *The concept of electronic record. In Preservation of the integrity of electronic records*, pages 9–22. Springer.

[11] Engeström, Y. 1987. Learning by expanding. An activity-theoretical approach to developmental research.

[12] Hill, G. 2004. An information-theoretic model of customer information quality. *Proceedings of the Decision Support Systems Conference*, Prato, Italy, July, pages 1–3. Citeseer.

[13] Klir, G. J. 2003. An update on generalized information theory. *Proceedings of ISIPTA 03*, pages 321–334.

[14] Shannon, C. E., and Weaver W. 1949. *The mathematical theory of information*.

[15] Star, S. L. 2010. *This is not a boundary object: Reflections on the origin of a concept. Science, Technology & Human Values*, 35(5):601–617.

[16] Star, S. L., and Griesemer, J. R. 1989. Institutional ecology, 'translations' and boundary objects: Amateurs and professionals in Berkeley's Museum of Vertebrate Zoology, 1907-39. *Social studies of science*, 19(3):387–420.

[17] Wang, R.Y., and Strong, D.M. 1996. Beyond accuracy: What data quality means to data consumers. *Journal of Management Information Systems*, 12(4):5–34, Spring.

[18] Rabinovici-Cohen, S., Factor, M., Naor, D., Ramati, L., Reshef, P., Ronen, S., Satran, J., and Giaretta, D. July/September 2008. Preservation DataStores,: New storage paradigm for preservation environments. *IBM Journal of Research and Development*, *Special Issue on Storage Technologies and Systems*, 52(4/5):389–399.

[19] Rabinovici-Cohen, S., Marberg, J., Nagin, K., and Pease, D., March 2013. PDS Cloud: Long term digital preservation in the cloud. In IC2E 2013: *Proceedings of the IEEE International Conference on Cloud Engineering*, San Francisco, CA.

[20] Jclouds. http://jclouds.org.

[21] Anderson, A., 2005. Hierarchical resource profile of XACML v2.0. DOI=http://docs.oasis-open.org/xacml/2.0/access_control-xacml-2.0-hier-profile-spec-os.pdf.

[22] Anderson, A., 2005. Multiple resource profile of XACML v2.0. DOI=http://docs.oasis-open.org/xacml/2.0/access_control-xacml-2.0-mult-profile-spec-os.pdf.

[23] Tarrant, D., and Carr, L. 2012. LDS3: Applying Digital Preservation Principals to Linked Data Systems. *Ninth International Conference on Digital Preservation (iPres2012)*, Toronto, CA.

[24] Unified digital format registry (UDFR). DOI=http://udfr.cdlib.org/.

Digital Preservation of a Process and its Application to e-Science Experiments

Stephan Strodl
SBA Research
Vienna, Austria
sstrodl@sba-research.org

Rudolf Mayer
SBA Research
Vienna, Austria
rmayer@sba-research.org

Gonçalo Antunes
INESC-ID
Lisbon, Portugal
goncalo.antunes@ist.utl.pt

Daniel Draws
SQS
Cologne, Germany
daniel.draws@sqs.com

Andreas Rauber
Vienna University of Technology
Vienna, Austria
rauber@ifs.tuwien.ac.at

ABSTRACT

The increase in computationally intensive science (called e-science) drives the need to make scientific processes available for the long term. The current approach is often to archive only the resulting publications, and at very most the data sets, of scientific experiments, which is insufficient in experimental and data intensive science. The preservation of scientific experiments and their results enables others to reproduce and verify the results as well as build on the result of earlier work. The TIMBUS projects aims at preserving processes for the long term. In this paper we present the process framework developed, and apply it to the preservation of a Music Classification evaluation process. This classification experiment represents a typical information retrieval process for classifying music into predefined categories, and evaluating the performance thereof. The paper describes and applies the process steps of the three phases of the TIMBUS approach: plan, preserve and redeploy.

Keywords

Digital Preservation, E-Science

1. INTRODUCTION

Digital preservation ensures the access to digital information objects over time. The main focus of research, so far, has targeted static digital objects such as text and multimedia documents. Recently, however, there is an increasing demand for preservation of dynamic objects and whole workflows and processes. The preservation of workflows and process is driven, besides others, by research institutions that run data intensive experiments. These experiments and their results need to be verifiable to others in the community. They need to be preserved as researchers need to be able to reproduce and build on top of earlier experiments to verify and expand on the results. Current practice in many disciplines is however often restricted to publishing results as a summary in scientific publications without detailed specification of the experiments. This is in some settings augmented by also making the data sets utilised available, but the lack of detailed information about the execution of the experiments, or the availability of the software employed, poses problems for the re-use in the long term. To avoid the loss of scientific results, work on digital preservation has thus expanded from a data centric perspective towards approaches to preserve the process to execute, render and analyse data.

Processes are increasingly supported by service oriented architectures, employing numerous services offered by different, external providers. These dependencies on third party services pose new challenges for the long term usability of processes. Software services are in general not designed for long term availability, as they rely on a number of technologies for execution, for example hardware, file formats, operating systems and other software libraries, which all face the risks of obsolescence. In the long run, the availability of today's technology cannot be guaranteed. The authentic functionality of processes in the long term can therefore be violated in terms of missing software services and outdated and unavailable technology.

The TIMBUS project[1] thus focuses on the preservation of (business) processes. The developed approaches and methods are domain independent and can be applied to different settings (e.g. business settings or E-Science domain). By analysing the execution context and identification of dependencies, the accessibility to processes and the supporting services is maintained over time. In this paper, we present the TIMBUS Preservation Process Framework, which specifies the process steps for the digital preservation of a process. The application of the preservation process is demonstrated on a use case process of a Music Classification experiment. This scientific process evaluates the performance of methods to classify music into sets of predefined genres, and is a typical task in Music Information Retrieval research.

[1] http://timbusproject.net

The remainder of this paper is organised as follows. Section 2 points out related work in the field of digital preservation with focus on holistic life cycle approaches. The Music Classification process is introduced in Section 3. The TIMBUS Preservation Process is explained in Section 4, showing the application of the music classification process. The paper concludes with a summary and outlook provided in Section 5.

2. RELATED WORK

Although digital preservation has been traditionally driven by memory institutions and the cultural heritage sector [18], it is increasingly recognized that it is a problem affecting all organizations that manage information over time, and as such it affects most of contemporary organizations where information systems provide important support to the business. Although the OAIS Reference Model [8] remains an important source of concepts to the field, it lacks directives and guidelines to address complex preservation scenarios with multiple business support systems and complex digital objects in place. In such scenarios, digital preservation requires a holistic view, acting as a combination of organisational and business aspects with system and technological aspects, so that all the contextual aspects surrounding a complex digital object can be captured and the objective of rendering it in the future in the same or in similar conditions can be attained.

With this holistic concern in mind, digital information life cycle models have been designed, of which the DCC Curation Life Cycle Model [6] and the SHAMAN Information Life Cycle [3] are noticeable examples. The DCC Curation Life Cycle Model elongates the traditional scope of preservation to include curation. It addresses two phases: a *Curation* phase, which might involve the creation of new information or the access and reuse of already existing information and its appraisal and selection; and a *Preservation* phase, which involves the ingestion of the information into the archive, the application of preservation actions, and the storing of that information. During the two phases, community watch and participation and preservation planning take place in order to keep descriptive metadata and representation information up to date. The SHAMAN Information Life Cycle, besides including the *Archival* phase already addressed by the OAIS model, suggests two additional pre-ingest phases and two additional post-access phases. The pre-ingest phases *Production* and *Assembly* aim at the capturing of the context of production of the object and its assembly into an information package, respectively. The post-access phases *Adoption* and *Use* concern the preparation of the retrieved package so that its information contents can be used.

This renewed understanding of preservation also creates the need for the development of new conceptual models that are able to synthesize this knowledge and make it re-applicable to different scenarios. The SHAMAN Reference Architecture [2] resulted from an infusion of knowledge in the digital preservation field and standards and best practices from the business and IT governance fields. It defines a set of preservation capabilities and their relationships and interaction with other organizational capabilities, so that its integration with the overall capabilities of an organization is facilitated. The overall objective is to promote the alignment between the preservation objectives of the organization, the organization's processes, and the existing technological infrastructure. Additionally, the work done on the CASPAR project on Preservation Networks [5] is a relevant reference on the capturing of the dependencies of complex digital objects through the usage of entity-relationship-like models, although business and organizational aspects are left out of it.

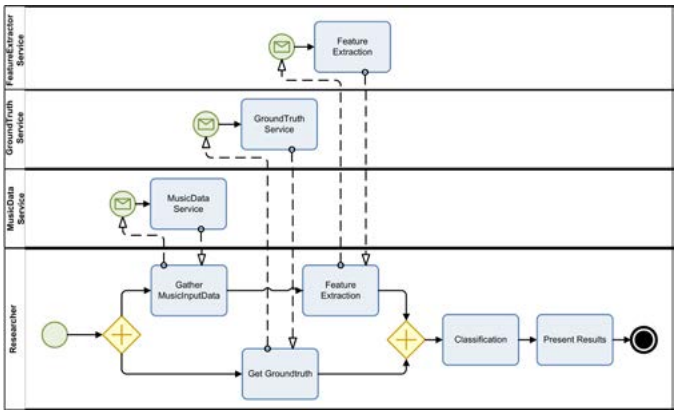


Figure 1: Music Classification experiment

Despite all the works referred to in this section, the preservation of business processes (in the form of their digital representation) along with the surrounding context needed for its long-term understandability is an innovative target being pursued by TIMBUS.

3. MUSIC CLASSIFICATION PROCESS

The process used in our case study is a scientific experiment in the domain of data mining, where the researcher evaluates the performance of an automatic classification of music into a set of predefined categories. This type of experiment is a standard scenario in Music Information Retrieval research, and is used with many slight variations in set-up for numerous evaluation settings, ranging from ad-hoc experiments to benchmark evaluations such as e.g. the MIREX genre classification or artist identification tasks [11].

The experiment involves several steps, which can partially be parallelised. First, music data is acquired from sources such as benchmark repositories or, in more complex settings, online content providers, and in the same time, genre assignments for the pieces of music are obtained from ground truth registries, frequently from websites such as Musicbrainz[2]. Tools are employed to extract numerical features describing certain characteristics of the audio files. In the case of the experimental set-up used for the case study, we employ an external Web service to extract such features. This forms the basis for learning a machine learning model using the WEKA machine learning software, which is finally employed to evaluate the prediction accuracy of genre labels for unknown music. The process is visualised using the BPMN notation in Figure 1.

[2]`http://musicbrainz.org/`

The process described above can be seen as prototypical from a range of e-Science processes, consisting both of external as well as locally available (intermediate) data, external Web services as well as locally installed software used in the processing of the workflow. In the implementation considered in this paper, it primarily consists of the following components:

- The Taverna workflow engine[3] is used to orchestrate the parallel execution and synchronisation of the process steps. Taverna further provides a scripting language based on Java that is employed for the above mentioned script tasks.
- A number of external services, all called from scripts or templates provided by Taverna, are employed:
  - The data source providing the music data is an archive with web interface, and can thus be obtained via HTTP requests.
  - The service offering the ground truth annotations, e.g. the assignment of a piece of music to a genre, is also obtained via HTTP.
  - The web service to extract features is a free service and similar to the one provided e.g. by Echonest[4]. In particular, we use a REST service that takes an MP3 file as input, and provides a vector of floating point values as descriptor.

  These services are provided by third parties, and their availability and similar function is thus not guaranteed in the future.

An illustration of the steps in this implementation of the process is given in Figure 2.

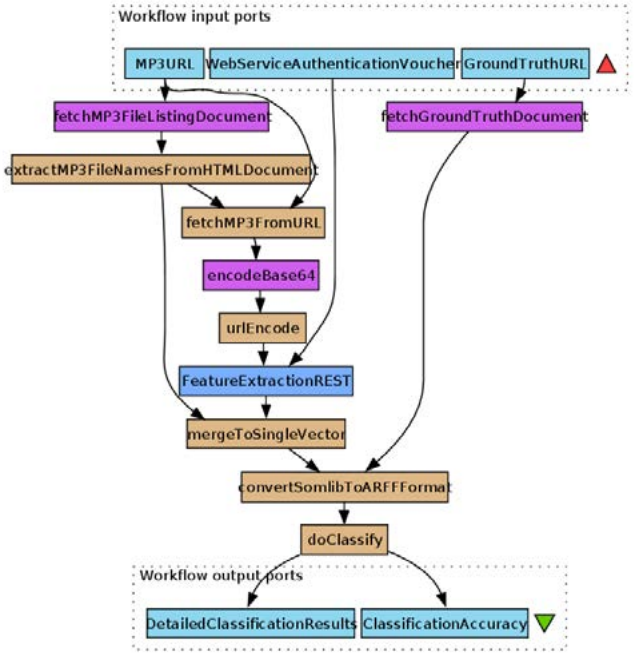


Figure 2: Music Classification experiment implemented in the Taverna workflow engine

4. TIMBUS PRESERVATION PROCESS

The TIMBUS Preservation Process to digital preserve business processes can be divided into three phases: plan, preserve and redeploy. The planning phase concerns the capture of the business process and its context. Risks of the business process are identified by reviewing contractual, policy and legal obligations. Driven from the risk management perspective, digital preservation is considered as a potential mitigation strategy. The assessment of preservation strategies identifies and evaluates different approaches to make the process available in the future. Figure 3 shows the TIMBUS process. Triggered by the risk management, the acquisition of the business process context and the *Assessment of Preservation Approaches* are executed in the planning phase (described in detail in Section 4.1). Within the preservation phase, presented in Section 4.2, the process data from the source environment are captured, preservation actions are executed and the data are prepared for archival storage. The redeployment phase, described in Section 4.3, specifies the re-initiating of the preserved process in a new environment at some point in the future. The fundamental concepts of the TIMBUS Preservation Process are presented in this paper, a detailed specification and description can be found in [16]. The process is domain independent and can be applied to different settings. In this paper we shown the application of the TIMBUS Preservation Process on the Music Classification process, which was introduced in Section 3.

[3] `http://www.taverna.org.uk`
[4] `http://the.echonest.com`

4.1 Planning phase

The planning phase is responsible to capture the process and its context and the assessment of suitable preservation approaches. As shown in Figure 3, the first step is the acquisition of the process context, followed by the risk assessment process. The risk assessment triggers the assessment of preservation approaches sub-process for identification, specification and evaluation of preservation strategies for the process.

4.1.1 Acquisition of the business process context

To successfully capture and archive the context of a business process, we have devised a context meta-model to systematically capture aspects of a process that are essential for its preservation and verification upon later re-execution [1].. This model is in the form of an OWL ontology, which enables checks for conformance and reasoning.

As the context of a process can involve a huge variety of different concepts, it is important to design a meta-model that is on the one hand generic, and on the other hand extensible to cover very specific aspects. We thus utilise a domain-independent ontology (DIO) that provides the generic core concepts, and domain-specific ontologies (DSOs) that are integrated and mapped to the DIO, and can refine its concepts. As the context of a process includes aspects on various different layers, the DIO is based on existing work in

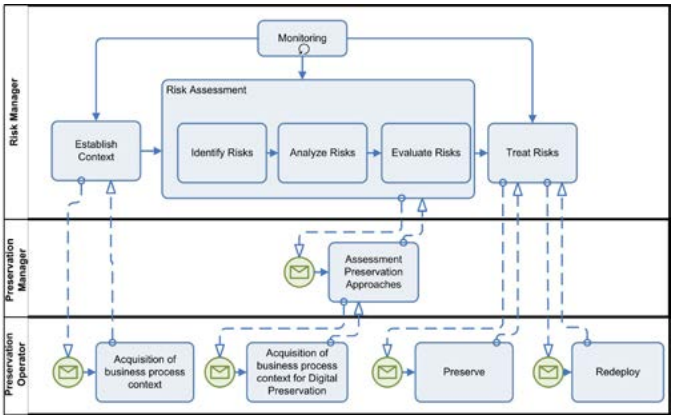


Figure 3: TIMBUS Process

enterprise architecture. Specifically, we adopted the Archimate [15] language, which provides a template to describe a business by around 30 different concepts on the business, application and technology layer.

We then further developed a number of domain specific ontologies that refine these concepts, including:

- Software licenses, based on The Software Ontology[5]
- Patents, based on the Patent Metadata Ontology (PMO), developed by the PATEXpert project[6]
- Software application dependencies, based on the CUDF, the Common Upgradeability Description Format[17]
- Digital preservation meta-data, based on the PREMIS data dictionary [13]

Some elements in these domain-specific ontologies are identified as sub-types of concepts defined in the domain-independent ontologies, and are mapped to these respective elements. This allows for a comprehensive description of the domain-specific aspects, while keeping the core ontology minimal.

The meta-model needs then to be instantiated for a specific use case. The context model can be further extended with other DSOs to define domain specific aspects of the processes. Some parts can be acquired automatically, such as the software dependencies on package-based operating systems such as Debian Linux, which also provides means to identify the licenses a certain package is distributed under. Other elements will have to be provided manually, for which we provide a graphical editor, implemented as a plugin to the Protégé ontology editor[7].

4.1.2 Risk Management

Risk management is a well establish field with the goal of defined prevention and control mechanism to address risks

[5] `http://theswo.sourceforge.net`
[6] `http://www.patexpert.org`
[7] `http://protege.stanford.edu`

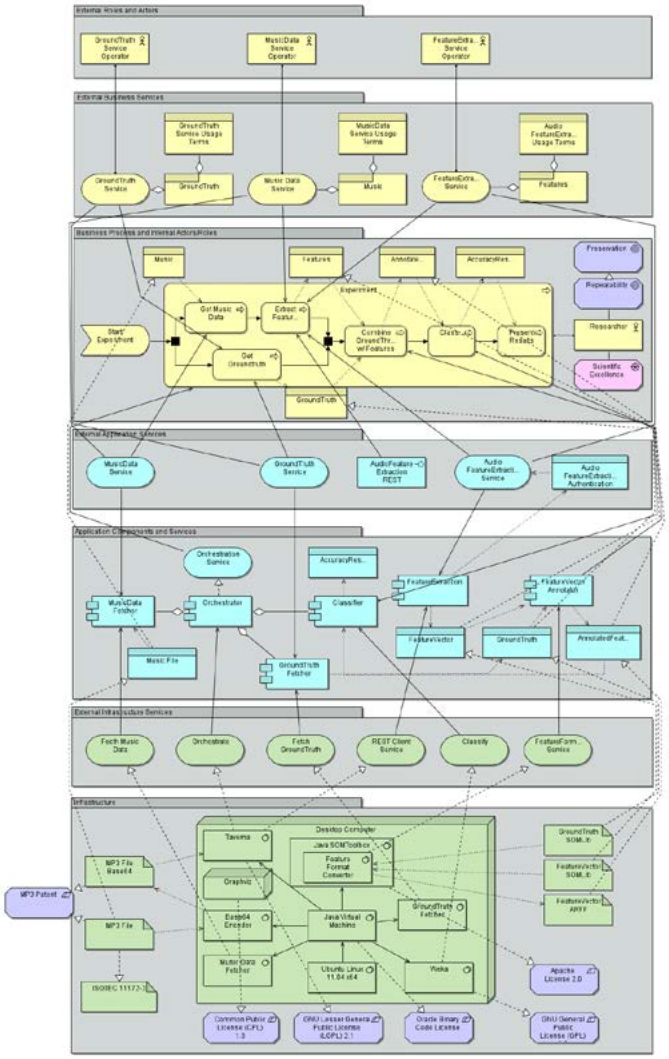


Figure 4: Context Model instance of the Music Classification experiment

related with assets and activities. Preservation can be seen as a potential method to mitigate risks, derived from the potential loss of information over time. The risk management process used in TIMBUS is based on the ISO 31000 standard [7]. Figure 3 shows the risk management process steps in the upper swim lane of the Risk Manager. TIMBUS defines the process-related interfaces to connect digital preservation with risk management. From the TIMBUS perspective, the risk associated to a process can act as a driver prompting its preservation as a way of mitigating the threats that endanger it. Risk management helps to identify and evaluate different risks in a structured and well defined manner. If information related risks have been identified and evaluated, different alternatives to preserve an adequate set of information have to be developed. The risk management process then decides what the best matching solution is. It triggers the preservation process, if a preservation alternative was assessed as a risk treatment.

For the Music Classification process the risk management is motivated by the goal to establish the institution as a sustainable excellence research center. According to the institutional polices, published scientific experiments and results need to be reproducible and verifiable in the future. Another goal of the institution is to foster the reuse and expand previous scientific work. This requires to share and reuse scientific experiments within the institution.

The use of external services represents a risk for the re-execution and verification of the Music Classification process, as the availability of the external services used cannot be guaranteed in the long term. Another risk identified is the lack of the documentation of the executed experiments. The results of experiments are published in conference paper, journals or reports, but these information are not sufficient to re-produce the experiment. Input data set, used software and parameter settings are often not specified in detail or not available any more. Moreover the technological dependencies of the experiment setting and a potential technical obsolesces were identified as potential risks for the Music Classification workflow.

A more detailed analysis of the risk assessment with respect to digital preservation for e-science processes is presented in [4]. After the risks are assessed, mitigation strategies are requested from the *Assessment of Preservation Approaches* process by the risk management as shown in Figure 3.

4.1.3 Assessment of Preservation Approaches

The *Assessment of Preservation Approaches* process is responsible for the identification and evaluation of different preservation approaches for the process. It starts with the refinement of the context model. In the first iteration the context model was created for the use by risk management. More detailed information about the technical implementation of the process is required for the planning of preservation strategies. The preservation requirements of the process are specified and documented for the evaluation and comparison of preservation approaches. The requirements specify the significant properties, describing functional and non-functional requirements of the process that need to be maintained over time. Redeployment scenarios support the specification of the significant properties regarding the preservation of artefact and execution of the process in terms of performance and behaviour. Different redeployment scenarios for future usage can be considered, e.g. execution of the original process with original data for confirmation of documented outcomes, execution of the original process with new data, or to modify parts of the process but using the original data e.g. for scientific workflows to evaluate improvement of new methods or models on the experiment results. For verification of the results from the Music Classification workflow, the original process needs to be re-executed with the original data from the executed experiments. Other preservation requirements can include amongst orders compliances to standards, institutional policies or legal obligations. The requirements are later used to evaluate and compare different preservation approaches.

The context model describes the implementation of the process. The primary goal is the preservation of the business logic of the process, not all implementation aspects are relevant for the future. Thus some implementation details can be abstracted and replaced to higher concepts. The abstraction of technical details can facilitate the preservation of the components (e.g. replacement through alternative implementations, use of emulators, or encapsulation). The level of abstraction and the aspects that can be generalised depend on the specific setting and the preservation requirements. As the abstraction causes loss of information, it is vital to ensure that no relevant information that is required in the future is lost during this step. An example for the music workflow is the operating system that can be generalised. As the workflow runs within the Taverna workflow engine, the underlying operating system does not represent a significant property of the experiment.

A process is an orchestration of tasks that are executed in a particular sequence, it can be complex, involving different services from various systems. For this reason, a combination of different preservation actions can be applied to preserve a process for the long term. Examples are virtualisation and emulation approaches for preserving functionality of services, and migration for documents.

A challenging task for the preservation of complex processes is the preservation of relationships and dependencies between components over time. Knowledge of the dependencies is important for maintaining the functionality of the components. Broken dependencies can prevent the redeployment of the process in the future. Examples are manifold, such as missing libraries for software execution, missing databases for data input, incompatible hardware for operating systems or missing credentials for encrypted data. The dependencies need to be considered whenever changes are applied to components. Modification of components for preservation purposes for example can have undesired side effects on other components. Examples are the migration of data into other formats that cannot be processed further by other software components, or the replacement of software components by new versions that offer different interfaces for interaction. Reasoning and queries based on the context model can help to identify decencies and further try to determine feasible preservation approaches [12].

While strategies for digital preservation, so far, mainly focus on data migration and emulation, the preservation of processes need further approaches, especially with respect to external dependencies. Different strategies can be used to maintain the significant properties of the process over time. Examples of strategies that support preservation and archival storage of processes are:

Metadata/Documentation
In order to maintain the usability, interpretability, accessibility and understandability of the process, additional metadata of its components are required. Understandability involves providing sufficient information so the component can be interpreted and understood in the future. Manual steps of the process that are not implemented by information systems require sufficient documentation and description for later redeployment. Furthermore logging and tracking functionalities of SW components (such as workflow engines) can be used to document the process execution and provide provenance information for the future [9].

Migration
Migration can be seen as the copying or conversion of digital objects from one technology to another. It is a widely adopted strategy for storage media and data formats. Besides that, the migration to alternative software services or components can be a vital approach for processes. For example in terms of licences, the use of alternative open source resources that provide the same functionality can be suitable strategy to overcome legal conflicts. Another aspect of migration can raise from the use of external services. As the availability of external services cannot be assured, a potential strategy is to transfer external services into the own system (in-housing). The strategy requires access to the implementation and data of the service as well as the licences and rights to operate the service. An example is cloud storage services that are operated by third parties.

Emulation
An emulator software mimics the behaviour and functionality of components, hardware or software. Emulation is a widely adopted strategy to preserve older computer platforms (e.g. video game console systems) and operating systems.

Virtualisation
Virtualisation (most common hardware virtualisation) has become a common business practice for server management. Virtualisation software provides a separation layer between the application services and the underlying hardware resources. The separation from actual hardware provides an abstraction layer of the physical environment, such as network, storage and display. It increases the robustness of virtual machines (VM) against changes of the underlying hardware. The virtualisation is a practical approach to capture complex systems to maintain the dependencies within VMs.

Mock-up of SW Services
A special problem for preservation represents the use of third party software services (e.g. Web services) within a process. A potential solution can be a mock-up of the services in form of a simulation of the original service. The basic principle is to intercept and record messages from the original system between process and service, which the simulation can then use to respond to request that have been captured previously. The approach is limited, as it can only be used for deterministic services (i.e. services for which the request and response pair always match, and which themselves are not dependent on any external state), and the mock-up can only respond to messages that have been recorded in the original system. For simple services and for the preservation of particular instances of a process, the mock up can provide a suitable solution if no other possibilities are given. An analysis of mock-up strategies for Web services, and recommendations to make Web services more resilient in general, can be found [10].

Software Escrow
Processes are often using proprietary and customised software application and services. The software is in many cases delivered as closed source to the customer that means the source code remains at the vendor and only the binaries of the software are delivered to the customer. From the preservation perspective, this scenario limits the potential preservation strategies for the software, as the software cannot be adapted to changes in the execution environment in the future. Software Escrow offers a mitigation as it places a trustable third party between the developer and the customer. All artefacts relevant to the software development are deposited at the escrow agent and released to the customer in case of predefined events (e.g. when the vendor goes out of business, or does not want to further maintain the software).

Different approaches can be used to preserve a process, using different strategies or tools. Each approach is specified in a *Process Preservation Plan*. The plan also defines procedures for capturing the process data and later redeploying and verifying the process. In order to preserve the process, the components and process data need to be captured from the source systems. The acquired data need to be in a consistent state that redeployment leads to a valid state of the process (e.g. all database transaction are closed). The redeployment procedure defines the execution of the preserved process in a new environment in the future. In order to ensure that the process is redeployed correctly, a verification and validation procedure is required. It defines measurement points to check that the redeployed process shows the same significant properties as the original process.

The proposed planes are evaluated against the previous specified preservation requirements. The evaluation includes the assessment whether the proposed models and procedures are complete and correct and that all significant properties of the process are preserved. In case the evaluation shows that relevant aspects of the process are missing or requirements are not fulfilled, a feedback loop of the *Assessment of Preservation Approaches* process allows the refinement of the context model or the preservation plan specification. Different preservation plans can be evaluated, and the evaluation results are submitted to the risk management for decision making. The impact of the different strategies on identified risks is assessed and the best matching solution is selected for treatment.

For the music workflow a combination of strategies was identified as most suitable preservation approach. The client side including the workflow, the workflow engine and the classification engine is captured in a virtual machine using Virtual Box[8]. As a underlying operating system Linux is used, because the licensing and activation methods of current Windows release can cause interferences in the future. The Music Classification workflow uses three external services, the music data, ground-truth and the feature extractor service. As the music and ground-truth data are free available, we can deploy the service on the client side. For the feature extractor we need another approach as we have no access to the implementation. In order to verify the experiments in the future, a mock up of the feature extraction service can be used to capture and replay the communication to the Web service of the process execution. Publications and documentation are migrated in standardized formats, PDF and Word documents are migrated to PDF/A by using Adobe Acrobe Distiller. The available software documentation in

[8] `https://www.virtualbox.org`

HTML format remains in the format.

4.2 Preservation phase

The acquisition and preservation procedures of the *Process Preservation Plan* are applied to the business process in the preservation phase. The software and data of the process are captured from the source environment. Preservation actions are executed and the process is prepared for archival storage. Validation and verification data are captured from the source system for redeployment. For the Music Classification process a sample input set and expected output can be used to validate the redeployment. Other measurement points can be logging information created from the workflow engine during the process execution that can be used for verification in the future.

For the preservation, a empty VM image is created and Ubuntu 13.04[9] is installed as operating system. The Taverna workflow engine is set up for the Music Classification workflow. The data source and ground-truth are migrated to local services by using Apache HTTP Server[10]. A mock-up software is installed to capture the traffic between the workflow engine and the external Web service of the feature extraction. The scientific experiments are executed in order to capture the responses from the Web services for the music files used. The implementation of the feature extractor cannot be preserved, but the behaviour of the service is documented through capturing of the traffic for later verification. A replay software that mimics the web services including the captured data set is installed on the VM system. Documentation and publications of the process and its component are stored within the VM. Viewer applications for the documents are installed as well. This strategies allows to bundle all required software and information in a single VM image. It reduces the technical dependencies for a re-execution of the process to a compatible VM player. The correct execution of the preservation phase is verified by test wise re-instantiation of the VM and the execution of the process. The validation and verification procedure is applied and the results can be compared to the original process. The preserved process needs to implement all signification properties of the original process as defined by the preservation requirements. In a last step of the preservation phase the process data are stored in the archive. In order to ensure that the process can be executed in the future, monitoring criteria have to be defined. The monitoring includes the external dependencies of the preserved process, e.g. technical requirements to redeploy. But also the requirements and polices of the organisation for the preservation of the process needs to be observed.

4.3 Redeployment phase

The redeployment phase defines the reactivation of a preserved process in a new environment at some point in time. The key characteristics of new environment are captured including the available technical components, organisational and legal aspects. A gap analysis between the requirements of the preserved process for redeployment and the available environment is performed. The technical infrastructure needs to be adjusted and prepared for the redeployment, different approaches can be used to overcome identified gaps, e.g. tools to emulate components, or migration of data formats. Required software and data are installed according to the redeployment procedure defined in the *Process Preservation Plan*. Tools and components for validation and verification are also set up in the new environment. As a final step the process can be re-executed and the taken measurements can be validated.

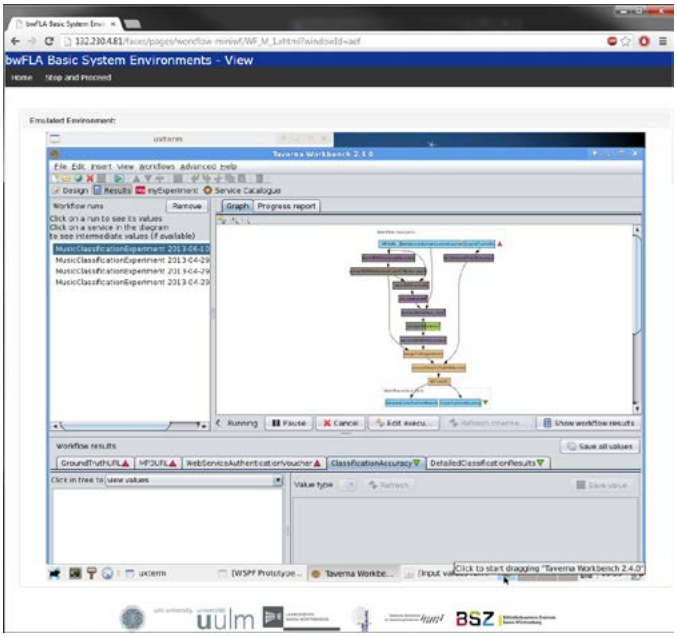


Figure 5: Emulation-as-a-Service of the Music Classification experiment

The archived Music Classification workflow requires a compatible VM player for the redeployment. While currently enough players for the format are available, the format needs constant monitoring to ensure the availability in the future. The use of external services for the redeployment can help to reduce investment and effort for the hosting institution. Emulation-as-a-Service can provide an interface to render preserved virtualised system. An example is the bwFLA project[11] that provides a web-based access to different rendering environment by using an emulation approach [14]. Figure 5 shows the Music Classification workflow that can be accessed via a Web browser interface and re-executed the captured process instances of the experiment. Captured validation and verification data can be used to check the correct behaviour of the redeployed process. Published results of the classifier can easily be verified while the reuse of the workflow is limited due to mock-up of the feature extraction. The mock-up service can only provide feature sets for music files that have been captured and preserved in the preservation phase. But experiments with new or modified classifiers can be done with the same music data set in the future, for example determining performance improvements or new classification approaches.

[9] `http://www.ubuntu.com/`
[10] `http://httpd.apache.org`
[11] `http://bw-fla.uni-freiburg.de`

5. CONCLUSION

This paper presents the TIMBUS Preservation Process to preserve processes for the long term. The process provides a guideline for the required steps to plan and perform the preservation and the later redeployment of processes. Driven from a risk management perspective, Digital Preservation is considered as mitigation strategy to address potential loss of information over time.

To preserve a process for the future, its influence factors and implementation need to be understood. Hence the context of the process is acquired, relevant aspects are identified that need to be maintained for the future. Potential preservation strategies are identified and tested considering the specific conditions, requirements, and goals of a setting. Different approaches can be combined to maintain the process over time. Processes are often implemented by using a service-oriented architecture implemented on a distributed infrastructure. The paper presents preservation approaches that address the specific needs of process preservation such as the preservation of external services. The redeployment reruns the archived process in a new environment at some time in the future. The process needs to be adapted according to the conditions of the new environment. The behaviour of the redeployed process is verified by comparing its measurements with measurements taken from the original process.

The application of the framework was presented in this paper by the preservation of a scientific workflow for music classification. The Taverna workflow engine was used to design and execute the Music Classification process. The process uses three external services that provide music files, ground-truth data and a feature extraction service. For the preservation of the process two of them were migrated to local alternative services implementing the same functionality. For the feature extraction a mock-up services was used to record the communication between the process and the Web service. The records captured can be used to mock up the service in the future and replay the communication. All software components used by the process were set up on a virtual machine. Documentation and publications about the scientific workflow were also stored on the VM. The result is an encapsulated process in a VM that can be archived. A potential redeployment strategy represents Emulation-as-a-Service where emulators for rendering environments are provided as Web service.

Acknowledgments

This work has been co-funded by COMET K1, FFG - Austrian Research Promotion Agency and by the TIMBUS project, co-funded by the European Union under the 7th Framework Programme (FP7/2007-2013) under grant agreement no. 269940. The authors are solely responsible for the content of this paper.

6. REFERENCES

[1] G. Antunes, A. Caetano, M. Bakhshandeh, R. Mayer, and J. Borbinha. Using ontologies for enterprise architecture model alignment. In *Proceedings of the 4th Workshop on Business and IT Alignment (BITA 2013)*, Poznan, Poland, June 19 2013.

[2] C. Becker, G. Antunes, J. Barateiro, and R. Vieira. A capability model for digital preservation - analyzing concerns, drivers, constraints, capabilities and maturities. In *Proceedings of the 8th Int. Conf. on Preservation of Digital Objects (iPRES 2011)*, 2011.

[3] H. Brocks, A. Kranstedt, G. Jaschke, and M. Hemmje. *Smart Information and Knowledge Management*, chapter Modeling Context for Digital Preservation, pages 197–226. Springer Berlin/Heidelberg, 2010.

[4] S. Canteiro and J. Barateiro. Risk assessment in digital preservation of e-science data and processes. In *Proceedings of the 8th International Conference on Preservation of Digital Objects (iPRES 2011)*, 2011.

[5] E. Conway, B. Matthews, D. Giaretta, S. Lambert, and M. Wilson. Managing risks in the preservation of research data with preservation network. *The International Journal of Digital Curation*, 7:3–15, 2012.

[6] S. Higgins. The DCC curation lifecycle model. *The International Journal of Digital Curation*, 3:134–140, 2008.

[7] ISO. *ISO 31000: 2009 Risk management – Principles and Guidelines.*

[8] ISO. *Space data and information transfer systems – Open archival information system – Reference model (ISO 14721:2003)*, 2003.

[9] R. Mayer, S. Proell, and A. Rauber. On the applicability of workflow management systems for the preservation of business processes. In *Proceedings of the 9th Int. Conf. on Digital Preservation (iPres 2012)*, pages 58–65, Toronto, Canada, October 1-5 2012.

[10] T. Miksa, R. Mayer, and A. Rauber. Ensuring sustainability of web services dependent processes. *International Journal of Computational Science and Engineering (IJCSE)*, 2013. Accepted for publication.

[11] Music Information Retrieval Evaluation eXchange (MIREX). Website. `http://www.music-ir.org/mirex`.

[12] M. A. Neumann, H. Miri, J. Thomson, G. Antunes, R. Mayer, and M. Beigl. Towards a decision support architecture for digital preservation of business processes. In *Proceedings of the 9th Int. Conf. on Digital Preservation (iPres 2012)*, 2012.

[13] PREMIS Editorial Committee. Premis data dictionary for preservation metadata. Technical report, March 2008.

[14] K. Rechert, D. von Suchodoletz, and I. Valizada. bwFLA – practical approach to functional access strategies. In *Proceedings of the 9th Int. Conf. on Preservation of Digital Objects (iPRES2012)*, 2012.

[15] The Open Group. *ArchiMate 2.0 Specification.* 2012.

[16] TIMBUS consortium. D4.6: Use Case Specific DP & Holistic Escrow. Technical report, 2013.

[17] R. Treinen and S. Zacchiroli. Description of the CUDF Format. Technical report, 2008. http://arxiv.org/abs/0811.3621.

[18] C. Webb. *Guidelines for the Preservation of Digital Heritage.* National Library of Australia, 2005.

Framework for Verification of Preserved and Redeployed Processes

Tomasz Miksa
Stefan Pröll
Rudolf Mayer
Stephan Strodl
Secure Business Austria
Vienna, Austria
{tmiksa, sproell, rmayer, sstrodl}@sba-research.org

Ricardo Vieira
José Barateiro
INESC-ID Information Systems Group & LNEC
Lisbon, Portugal
{rjcv, jose.barateiro}@ist.utl.pt

Andreas Rauber
Vienna University of Technology & Secure Business Austria
Vienna, Austria
rauber@ifs.tuwien.ac.at

ABSTRACT

Preserving processes requires not only the identification of all process components, but also the interception of all interactions of the process with the external influencers. In order to verify if the collected data is sufficient for the purpose of redeployment, as well as to verify that the redeployed process performs according to expectations, a framework for verification is needed. This paper presents a framework for verification of preserved and redeployed processes. We demonstrate the applicability of the framework on an use case from the eScience domain. The preservation and the redeployment of the eScience process is tested by migrating it to substantially different environments.

1. INTRODUCTION

Traditionally, research in the area of digital preservation deals with preservation of static information like documents, scans, and other kinds of data. The long term preservation of entire systems and processes was not in the centre of attention. Addressing this new challenge requires advanced methods and processes which ensure that the process context is described adequately. This includes the collection of sufficient information of all involved components, which enables future redeployment. No matter how well-engineered the process for preservation of processes is, it cannot guarantee that all necessary information required to run the process was recorded. Given the complexity of preserving entire systems and processes, we thus need to derive means for reliably verifying whether a process being re-deployed performs correctly according to preservation goals. We need to ensure that not only sufficient information is collected during planning and preserving of the process, but also to confirm that the redeployed process performs according to the expectations of the redeployment scenario.

The verification of redeployed processes is a complex task which may vary in its form due to several factors: the way the processes are specified, the drivers for their preservation, the preservation strategies applied; the reasons for the redeployment, the redeployment environments, etc. However, regardless of these differences, all processes must be verified for measuring the success of the redeployment. Otherwise, there is no guarantee that the process running in the redeployed environment is the one which was meant to be redeployed. Such evidence is crucial in litigation cases when the correctness of the original process, executed at some time in the past, could be questioned, and the only way to check this is to re-run the original process. In such cases, the method for verification of redeployed processes should provide irrefutable evidence that the redeployed process is behaving exactly the same way as the original. On other perspective, in the domain of eScience [7] and Research Infrastructures [11], where scientists make scientific discoveries by creating and constantly improving the processes for transforming Big Data [10], the verification of redeployed processes is essential. It enables researchers to verify their results, or apply previously used models on new data.

In this paper, we present the VFramework which is a framework for verification of preserved process. It is a refinement of a conceptual framework presented in [6]. It consists of 7 steps that describe the key actions which have to be performed in order to verify any kind of redeployed process. The VFramework can be applied not only to fully redeployed processes but is also capable of evaluating partial redeployments. Moreover, the VFramework can also verify both identical and re-engineered processes. We present an application of the VFramework for verification of a redeployment of an eScience process in the domain of sensor data analysis, which was extracted from its original environment and was redeployed in various environments different from the original one. The VFramework was applied to assess these redeployments.

The paper is organized as follows. Section 2 presents the state of the art. In Section 3 the steps and requirements set to the VFramework are described. Section 4 describes the application of the VFramework to the use case. We provide conclusions and future work in Section 5.

2. STATE OF THE ART

In [6] a conceptual framework for evaluation of emulation results was presented. It was demonstrated in [5], that the framework can be successfully applied to evaluate the conformance and performance quality of applications and processes redeployed in an emulator. This was demonstrated on case studies in which the framework was used to evaluate the emulation of a video game and an accounting program. The VFramework presented in this paper is a refinement of that framework for complex, potentially distributed processes. It provides detailed specification of actions which have to be performed for verification of redeployed processes.

In ISO 12207 [2] the life cycle processes for systems and software were defined. "It contains processes, activities, and tasks that are to be applied during the acquisition of a software product or service and during the supply, development, operation, maintenance and disposal of software products" [2]. It does not consider the redeployment as a part of the life cycle and hence provides no guidance for the scenario considered in this paper. The standard defines also the Software Specific Processes and lists actions which are needed for the Software Verification Process and the Software Validation Process. However, these processes belong to the Software Support Process category which assists the software implementation process. As a consequence, these processes are highly coupled with the software development, what is not in the scope of our investigations. Summing up, ISO 12207 does not specify a process for verification of redeployed software processes as presented in this paper.

The IEEE 1012 standard [1] specifies a process for software verification and validation. This process addresses the following software life cycle processes: acquisition, supply, development, operation and maintenance. It is compatible with ISO 12207. It defines tasks, required inputs and outputs to conduct verification and validation (V&V) of the software at all aforementioned life cycle processes. The V&V process for the maintenance process considers migrations to other environments. This overlaps with some of the requirements we set to the framework for verification of redeployed processes (see Section 3), i.e. the system is migrated to the other platform when the original system is still available. However, it does not consider the situation when the system or the process is disposed, deposited and redeployed after some time. Furthermore, the standard specifies only a high level list of activities applicable in several maintenance scenarios which are rather focused on verification and validation of the activities performed to keep the system running (e.g. system updates, bug fixing, enhancements to the functionality), rather than on digital preservation scenarios. The VFramework proposed in this paper provides more detailed guidance and can be applied to a broader range of digital preservation scenarios.

3. VFRAMEWORK

The VFramework was created to verify that a redeployed process performs according to expectations. There were two main requirements set to the framework.

Firstly, the framework has to be independent of the situation in which different digital preservation actions were applied to the full process or to different parts of the process. In such situations some of the process' parts may be substituted, re-engineered, emulated, migrated, etc. As a result, the redeployed process which is to be verified is not necessarily an exact copy of the original process. The framework has to be capable of verifying the execution of similar processes or their parts. By similarity of processes we mean a situation in which the functionality or characteristics of the process have been altered, but the deviation is either desired (e.g. faster computation) or acceptable (e.g. some functionality is limited but for the purpose of redeployment it is not required). Such situations may be an inevitable side effect of the digital preservation actions or a consequence of deliberate actions (e.g. improved implementation of the process). The framework has to support such situations regardless of its origin, and be capable of evaluating full and partial redeployments of processes.

Secondly, due to the high variety of the nature and implementation of the processes and a wide range of potential user requirements that had to be considered, the framework has to be flexible to cover all these requirements and settings. Therefore it has to remain at a relatively high level of abstraction and be customizable for the concrete processes which are going to be preserved. The guidance on customization has to be provided by the framework in order to achieve the comprehensiveness of the process verification.

The VFramework is depicted in Figure 1 and it consists of two sequences of actions. The first one (depicted in blue) is performed in the original environment. The results of the execution of each of the steps of this sequence are stored into the VPlan. The VPlan is a machine readable document in which all of the information about the original environment is kept. The second sequence (depicted in green) is performed in the redeployment environment. The necessary information for completion of each of the steps is obtained from the VPlan.

Original environment denotes the system in which the process, which is going to be preserved, is deployed and operates. The redeployment environment is the system in which the process will be installed once the decision to redeploy the preserved process is taken. It is very likely, that the redeployment will take place at some distant time in the future, when the original platform does not exist anymore and the process may need to be re-engineered to fit it into the new system.

Apart from descriptive metadata, the VFramework uses two kinds of data: verification data and redeployment performance data. The verification data is collected during the execution of the process in the original environment. It provides information on details of the execution of process instances, focusing on measuring significant properties. Interactions with external components have to be stored as well. For this purpose, external interaction data being part of verification data is collected. This external interaction data represents a record of all interactions of the process with external components during the execution of a specific process instance in a scenario to be used for verification. This data is reapplied in the redeployment environment to ensure determinism, by recreating the same external interactions. The redeployment performance data is collected during the

execution of the process in the redeployment environment. It provides information on details of the execution of the process instances, focusing on measuring significant properties. It is used for comparison with verification data to assess the redeployment. The steps of the framework are described below.

1. Describe the original environment The aim of this step is to describe the process and document its context by identifying environment dependencies in which the process is deployed. As motivation for the preservation of the process, considered redeployment scenarios and a set of example instances to be used for verification are determined. This corresponds largely to steps 1-3 of the "Define Requirements" phase in preservation planning [4], with the first step being subdivided into two more fine-grained steps.

1.1 Describe the process The information should describe the process itself but also the context in which the original process operated. A detailed description of not only software and hardware requirements, but also legal aspects is needed. Such information can be provided in multiple forms. One of them could be the context model [8], which is an ontology based model for description of processes and their dependencies.

1.2 Define set of potential redeployment scenarios The purpose of the redeployment has to be defined. This information has significant impact on the process of verification, because it impacts the type of measurements and the results they are supposed to fulfil. For example, different requirements are set to the process which is supposed to be an exact copy of the original process redeployed for a purpose of litigation case when the correctness of the original process has to be proven and therefore the redeployed process is verified for being identical. Different requirements are set to the eScience process which is redeployed with some of its parts substituted with components of the same functionality but improved quality (e.g. faster computation, more accurate results, etc.). In such cases some of the measurements may be ignored or interpreted differently, e.g. accuracy of results should not be worse than the original, but does not need to be exact. Verification focuses in this case on ensuring the functionality is achieved, but the significant properties related to part where the changes were introduced should be treated differently.

1.3 Select process instances to be used for verification Process may have several execution paths and therefore instances of the same process may vary considerably. In this step, the instances of the process which will be used for verification are selected. They have to be chosen according to the considered redeployment scenarios. The instances selected at this step will be used to collect both verification data from the original environment, as well as the performance redeployment data. The description of selected instances should specify in a comprehensive way all actions which were performed when running the process. These could be depicted by sequence diagrams, activity diagrams, use case diagrams, textual description, etc. The way it is specified depends on the level of automation of the process, e.g. if it is a manual process or formally specified in BPMN or executed within a workflow engine. Furthermore, the values of all parameters and input values must be documented.

1.4 Identify significant properties to be preserved The significant properties which have to be preserved and then evaluated have to be specified. They can either be collected at this step or obtained from preceding activities, e.g. preservation planning. However, regardless of the source, it is important that the significant properties reflect both functional and non-functional requirements of the process. It is important to determine which significant states of the object are to be measured as the significant property. These significant states could be: target state, continuous stream or series of states.

2. Prepare system for preservation The aim of this step is to identify the interactions of the process, i.e. all inputs and outputs of the process, but also configurations of process parameters, as well as influences of other components sharing the process environment or used indirectly by process components. This information is needed in order to ensure deterministic execution of the process and thus ensure reliable assessment. The steps should be conducted in view of redeployment scenarios and significant properties defined for the process.

2.1 Determine process boundaries The process boundary specifies which elements belong to the process and which elements belong to the external environment in which the process operates. It is possible to define different process boundaries depending on the scenarios for redeployment. For example, if the scenario assumes redeployment of only a part of the process which will be fitted into another process, then only the redeployed parts of the original process are within the boundary. However, there may be a second scenario in which the full process is redeployed, then a second boundary has to be defined which covers the entire process. Boundaries may also be influenced by the degree of control one can exert on specific components (e.g. external web services) and their importance for redeployment as well as their stability. In all cases, the description should ensure that the process boundaries are specified clearly, i.e. a distinction between elements which are part of the process to be preserved and which are external services with which the process exchanges data has to be made. This is particularly important in case of distributed processes which are using the Service Oriented Architecture for their implementation, or those deployed in the Cloud.

2.2 Determine external interactions For each of the specified boundaries the external interactions have to be identified. External interactions denote situations in which elements within the process boundary interact with elements from outside of the boundary. External interactions may be critical for the correct execution of the entire process, because any changes in the external components may cause changes in process execution. For example, the web service which provides data for one of the process steps may change (change of interface, implementation of algorithms, etc.) or become unavailable [9]. As a result, the process can perform differently (providing different outputs) or cannot run anymore. Another example could be encryption and the necessity to access an authentication service. When the certificate is not available anymore, then the communication

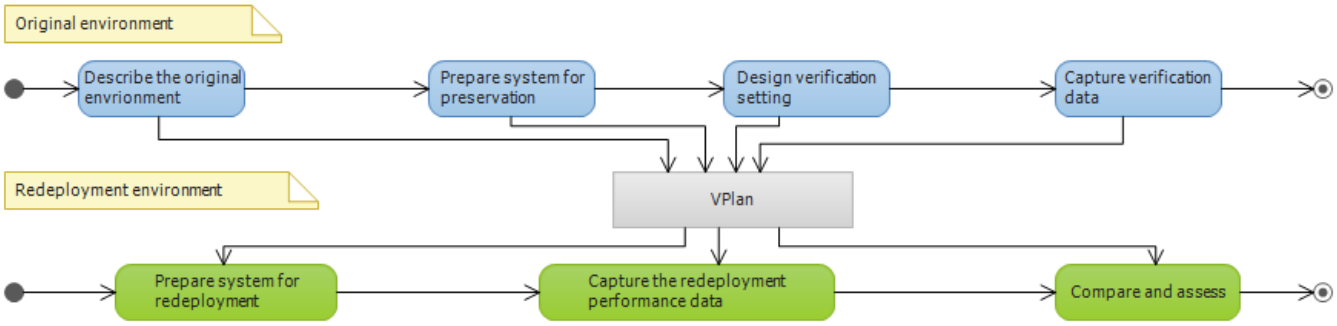


Figure 1: VFramework - framework for verification and validation of preserved business processes

cannot take place unless the authentication is removed (if the redeployment scenario allows this).

Special attention has to be paid to indirect external interactions and consequences for the process which might not always be visible at the first sight. For example the operating system if not included within the process boundary, its version and all system updates may alter the execution of the process. For all the requirements which focus on the visual presentation, the installed fonts, appearance settings, colour schemes of the system may be such influencers. Other digital objects which coexist in the system may also have impact. For example, processes running in the background (e.g. virus scan software, remote desktop software) can significantly affect the performance of a system. Moreover, other processes may share common data with the examined process and may modify the data that may result in the non-deterministic execution of the analysed process. Furthermore, all user or system I/O (e.g. keyboard, network, specific hardware components such as system clock, etc.) that are outside the process boundaries need to be identified.

2.3 Determine internal interactions The process may consist of several components which have their own settings. All these settings must be determined at this step. Furthermore, some of the process components depend on further software tools or libraries which may vary in version or settings. Some examples of these could be: virtual machines, database software, libraries, software device drivers, fonts, codecs, etc. The detected versions of components have to be verified to detect if the versions have not been modified or customized. If some of them were modified (e.g. modified config files) and this has an impact on the process, then they have to be preserved as well. Besides the software dependencies, the underlying hardware has to be considered when searching for potential internal interactions. The process may depend on some proprietary and unique hardware equipment or the underlying hardware may have some specific implementations of algorithms affecting the results obtained in the process. For example, some of the hardware bugs may affect the results delivered by the process. These results will only be achievable on a particular hardware platform (e.g. well-known Pentium FDIV bug had an impact on the results of floating point calculations, and therefore could alter the results of the whole process upon correct redeployment).

2.4 Ensure deterministic behaviour To allow verification of redeployment we need to ensure that a process performs deterministic. Thus, all interaction identified in 2.2 and 2.3 need to be verified for completeness to ensure deterministic re-execution. If this is not possible within the generic process, adaptations have to be made specifically for verification. If the determinism cannot be ensured, the verification of processes is not very likely to be possible. The investigation of determinism of the process should be conducted in view of considered redeployment purposes. In some settings, some of the non-deterministic influencers are affecting measures which are not important for the purpose of the redeployment. For example, when the exact execution speed is not considered a significant property, then all of the non-deterministic influencers regarding this particular criterion do not have to be considered.

When one of the process steps exchanges data with some third party component (external interaction), the communication can be recorded and replayed in the redeployment environment. If the process depends on the component which affects determinism of the process, it may be possible to substitute the component with a mock-up which does not have this deficiency. An example of such a solution for web services can be found in [9], if one of the steps of the non-deterministic process depends on a random number generator, then it may be substituted with a mock-up which always provides the same sequence of values as the one recorded and thus the process becomes deterministic. Of course, such changes to the process must be documented and possibly reverted after the verification process is finished in the actual redeployment, but for the purpose of verification they should be present.

3. Design verification setting The aim of this step is to identify the measurement points of the process, specify metrics used to assess quality of preservation actions and couple them with thresholds which are used as criteria for the assessment. The measurement points can be defined as points of the process where data enabling reasoning about correctness of the process execution is collected. The investigation should be conducted in the view of redeployment scenarios and significant properties defined for the process.

3.1 Specify measurement points Measurement points for both internal and external interactions must be described unambiguously and precisely, because the given value can

be measured in different ways and in different parts of the process and therefore not always the same values may be obtained. For example, the output of a process that transforms some images into PNG files is selected as a measurement point. This seems to be a clear requirement but without explicit definition of what is exactly measured the results may vary, because the bit streams which write the PNG file to the disk can be compared on the fly or the files already written to the disk can be opened and analysed by image recognition algorithms. In the first case, different libraries may have been used to transform the image (e.g. library was replaced in the redeployment) and as a result the outputs may be different at the bit level, while in the case of image recognition algorithms the images may turn out to be identical. Both approaches are valid and can be used. As the example shows, the choice of the measurement point depends on the requirements and intentions of the future redeployment. We thus need to identify, for each significant property of the process, on which level these must be captured. According to [6], the core levels are (1) bit level file storage, (2) the rendering of an internal state in a the system memory, (3) memory of an output device (e.g. video card memory (virtualized or real)), (4) port communication (e.g. VGA port, network interface, audio port) or (5) the actual output device (screen, speakers, actuator). If the verification aims to check if the rendering algorithms are exactly the same, then the bit comparison seems to be a better measurement point. But if it is allowed to modify the process and only the final visible product needs to be verified, then the second approach should be selected. It may be advisable to take measurements at multiple measurement points and collect the data for all of them. The choice of the measurement point which is most accurate for the redeployment environment will be left to the person redeploying the process who is aware of the reasons and requirements set to the redeployed process. While measurement points will usually relate to external interactions (e.g. result storage, communication with user or external system), internal interactions within process may be useful to capture for partial redeployments, to allow application and verification of a wider range of preservation actions (such as component replacement) and to allow more flexible redefinition of the boundaries identified in step 2.1.

3.2 Specify metrics for preservation quality comparison The significant properties which were selected in the first step have to be decomposed from high level significant properties into tangible and measurable metrics which can be measured and identified directly in the process. A wide range of techniques can be used for decomposition. Especially techniques stemming from requirements engineering may be particularly useful in this step, e.g. goal modelling [12], GQM method [3], etc. It is also advisable to specify metrics which can identify what the process should not do. In many cases it is easier and quicker to identify the forbidden behaviour or an incorrect state of the process. Then the redeployment can be rejected without a necessity of checking other metrics.

Having defined the metrics, the target values are assigned. These values will be used as the criteria for the assessment. They have to be specified in view of considered purposes of the redeployment. This information has significant impact on the process of verification, because it impacts the importance of available metrics and results they are supposed to achieve. Target values itself can be specified in different ways, e.g. metric A equals Y, metric B is maximum 120% of the original value, etc.

3.3 Aim for automated measurements capture When the VFramework is applied during planning of the preservation activities and different preservation scenarios and activities are considered, the possibility to automate measurements decreases the time needed for evaluation of alternative preservation strategies. This has lower importance when the VFramework is used during the preservation phase and redeployment phase, when the preservation strategies are already defined. Regardless of the phase, automation of measurements eases the process of verification.

4. Capture verification data This step has two main tasks. Firstly, to configure the capturing environment for collection of verification data. Secondly, to collect the verification data while the process is monitored by tools which trace process interactions.

4.1 Prepare system for capturing In this step the capture environment is configured. Either a clean environment is created in which the process is deployed, or an existing instance of an operational system is used directly.

4.2 Prepare data capture tools Tools for capturing external interactions, as well as verification data are introduced to the capture environment in the next two steps.

4.2.1 Set up tools for capturing external interactions Tools which will intercept external interactions of the process are installed in the capture environment. The captured information will be used to ensure deterministic execution of sample process instances (step 1) in the redeployment environment.

4.2.2 Set up tools for capturing verification data Tools which collect data in previously specified measurement points are installed in the capture environment. The captured information will be used to evaluate performance of the redeployed process.

4.3 Run the process and capture data When the capture environment has been configured and the tools for capturing data are in place, the instances of the process, which were identified in the first step, are executed. The data is being collected during and after the execution of the process.

4.4 Verify validity of captured data Once the execution of process instances has finished, the recorded data is verified for its correctness. This could be either manual or automatic action, which checks if all the measurements were stored correctly, e.g. if the log files are not empty. If all the data is correct then it is stored into the VPlan.

5. Prepare system for redeployment This is the first step performed in the redeployment environment. This step has three main objectives. Firstly, to configure the redeployment environment for collection of redeployment performance data. Secondly, to redeploy the process in the new environment. Thirdly, to execute process instances.

5.1 Prepare redeployment environment The environment in which the process will be redeployed has to be selected. Tools which ensure determinism during execution of the process, as well as the tools used for data collection have to be installed.

5.1.1 Set up redeployment system Either it will be a clean system or a system in which some other processes already exist. This depends on the purpose of the redeployment. If the process is run in an environment shared by other process an analysis of possible external interactions has to be conducted in order to ensure that the determinism of the redeployed process is not affected by the new environment.

5.1.2 Set up external interactions replay to ensure determinism The external interactions data is used in this step to recreate the interactions of the system. Tools which allow replaying of this data have to be installed in the redeployment environment.

5.1.3 Set up data capture tools Similarly to the step 4.2, the tools which extract redeployment performance data are installed in the redeployment environment. These tools will collect data needed for verification of the redeployed process at the predefined measurement points.

5.2 Redeploy preserved process The preserved process is redeployed in this step. Required adjustments to run the process in the new environment are done and the instances of the process which were used in the original environment are executed.

5.2.1 Identify required preservation actions to enable redeployment The aim of this step is to ensure that the process becomes operational in the new environment and that all of the instances of the process defined in the first step can be executed.

It is very likely that the preserved process will have to be re-engineered in order to be fitted into the new environment. For example, in the given environment a certain library responsible for encrypted communication with a web service cannot be used. However, a substitute library which allows to communicate with a web service with a different encryption mechanism might be available. Then such substitution has to be made in order to make the process operational (only if the redeployment scenario does not exclude such an action). In this step all kinds of preservation actions such as replacing a library with another one, cross-compiling code, migrating a file, putting an additional wrapper around the component, etc. may be applied.

5.2.2 Re-run the set of process instances Process instances, which were defined in the first step and executed in the original environment to collect verification data, are executed in this step in order to create redeployment performance data. The execution is controlled by the tools which ensure determinism of the process.

6. Capture redeployment performance data The aim of this step is to collect the redeployment performance data from the new system and verify if the data collection conditions were fulfilled.

6.1 Collect redeployment performance data The redeployment performance data is recorded by the tools which are monitoring the execution of process instances. All this data is collected and will be used for comparison with the verification data.

6.2 Verify validity of captured data Before the data can be used for comparison, its validity and fulfilment of assumed level of determinism of the environment needs to be checked.

6.2.1 Verify if required level of determinism was reached Results have to be analysed regarding the required level of determinism in the environment. If it was possible to ensure it and the tools which were introduced for this purpose in the step 5 performed its task correctly then the requirements are fulfilled. Otherwise, the procedure has to be repeated starting from the step 5 and new ways of ensuring deterministic execution of the process have to be introduced.

6.2.2 Verify correctness of capture data Similarly to the step 4.4, the collected redeployment performance data needs to be verified before it can be used for further analysis. This could be either manual or automatic action which checks if all the measurement were stored correctly, e.g. if the log files are not empty.

7. Compare and assess The comparison of significant properties measured in both environments is conducted in this step. The comparison is described in a report and a decision about fulfilment of redeployment purposes is made.

7.1 Compare redeployment performance data and verification data In this step the comparison between verification data and redeployment performance data is conducted. The comparison has to be done by contrasting the data collected at each of the measurement points of the original process with the data collected at each of the measurement points of the redeployed process. Due to the changes which might have been introduced to the process, some of the measurement points may not be available. If so, the comparison is either omitted or another corresponding point is used.

7.2 Conduct preservation quality comparison The metrics which were specified in Step 3.2 are calculated for the redeployed process. These metrics allow to assess the quality of preservation actions. These metrics are always interpreted depending on the redeployment scenario, because they may have different target values depending on the scenario. In some scenarios the specific functional or non-functional metric must be fulfilled, while in the other scenario it is not a requirement.

7.3 Provide summary report A report summarising the comparison is created. The report is supposed to deliver credible information about the state of the redeployed process, measurements made, metrics and their expected values and any alterations detected which are not compliant with the purpose of the redeployment.

7.4 Make the final decision The final decision is made by the auditor who knows the reason for the redeployment

and using the report can make a credible decision.

7.5 If positive, remove tools used for verification If the process is positively evaluated, then the tools for ensuring determinism are removed from the environment, unless they are needed for the redeployment. The original implementations or substitute services providing the full functionality are used instead. Similarly the tools for data collection can be removed from the environment.

4. VFRAMEWORK EVALUATION

In this section we test the applicability of the framework to an eScience use case. Section 4.1 provides details on the use case, Section 4.2 explains how the VFramework was executed to verify the process migration from Windows to Linux.

4.1 Use case description

The use case provider stems from the domain of civil engineering. It owns and maintains a system for supporting the process of acquiring and managing data captured from sensors installed in dams for monitoring the structures. The experts working for this institution execute many processes which are used for the structural monitoring through sensor networks to determine the actual structural state, managing visual and technical inspections to detect or analyse potential anomalies, and physical and mathematical models to estimate the structural behaviour. They also use data analysis tools such as tabular and chart reports and graphical representation of geo-referenced information. In fact, knowing the past structural behaviour is the best tool to perform complex analysis and make correct predictions about the state of the dams. In case of any anomaly or emergency, the sensor data may need to be reprocessed or reanalysed to look for mistakes in the original processes. Therefore being able to rerun the processes using either the original data or the new data and parameters is a crucial requirement for this organisation. Digital preservation of processes was selected as a strategy to address this requirement.

The process which is used for testing the applicability of the VFramework is depicted in Figure 2. This process is run by scientists who use their desktop workstations with Windows 7 as the operating system. The process consists of 5 steps which fetch the sensor data from an external web service (*Get Data Files*) download an R (*Get R script*)and TEX (*Get Tex script*) scripts for processing and compiling the data, generate PNG and TEX files (*Generate Plots*) which are finally compiled into a PDF report (*Generate Report*).

4.2 VFramework application

1. Describe the original environment The first step consists of sub steps which describe the process in detail, choose its significant properties and process instance used for data collection, as well as specify potential redeployment scenarios.

1.1 Describe the process We have described the purpose of the process, identified the users of the process and documented the components which build the process. We have used tools for detecting software dependencies and documentation provided by the owner of the process.

1.2 Define set of potential redeployment scenarios In this step we have defined two potential redeployment scenarios. Scenario 1 assumes that the process will be redeployed in order to be fully operational. It assumes that the external communications (e.g. web service) will be available. Scenario 2 assumes that the process will be redeployed in order to confirm that the values and plots presented in the scientific paper were obtained from a cited data set. The data set which was used for processing will be provided and there will be no need for communication with the web service. Further steps of this framework are always performed in view of requirements of these scenarios.

1.3 Select process instances to be used for verification For each scenario we have selected 10 instances which were differing in the configuration of parameters. For scenario 1, parameters for fetching data from a web service were randomly altered. For scenario 2, 10 data sets from 10 different locations were used.

1.4 Identify significant properties to be preserved We have conducted interviews with the owner of the process in order to collect the list of significant properties for each of the considered scenarios. We have collected both functional requirements, e.g. the system must be able to generate sensor data for quantitative interpretation, and non-functional, e.g. the system provides correct results. The significant properties were grouped by the scenario for which they are important (sometimes both).

2. Prepare system for preservation In this set of steps we have identified the process boundaries and described its interactions. Our analysis also included the determination of non-deterministic influencers and strategies for their mitigation.

2.1 Determine process boundaries There are two scenarios of redeployment considered. In the first one, which assumes that the process is redeployed in order to be fully operational, the presence of the web service is assumed. Therefore we put the web service outside of the process boundary. In case of the second scenario, when the redeployment is done for the purpose of validation of experiment's results, we exclude first three steps of the process (*Get Data Files*, *Get R script*, *Get Tex script*) from the process boundary. Data used in the original experiment will be applied to the process directly. In both cases the scripts which are the implementation of process steps: *Generate Plots*, *Generate Report* are within the process boundary. The operating system and the software required to run the scripts is not included as part of the process.

2.2 Determine external interactions In case of the scenario 1 the process transforms the data obtained from a web service. This is identified as an external interaction. In both redeployment scenarios, steps of the process are executed manually by executing commands using the keyboard. Therefore the input from the keyboard is another type of external interaction. Finally, the files produced as the output of the process are displayed on the LCD screen in the form of a PDF document. The visual presentation of the results on the screen is also identified as an external interaction.

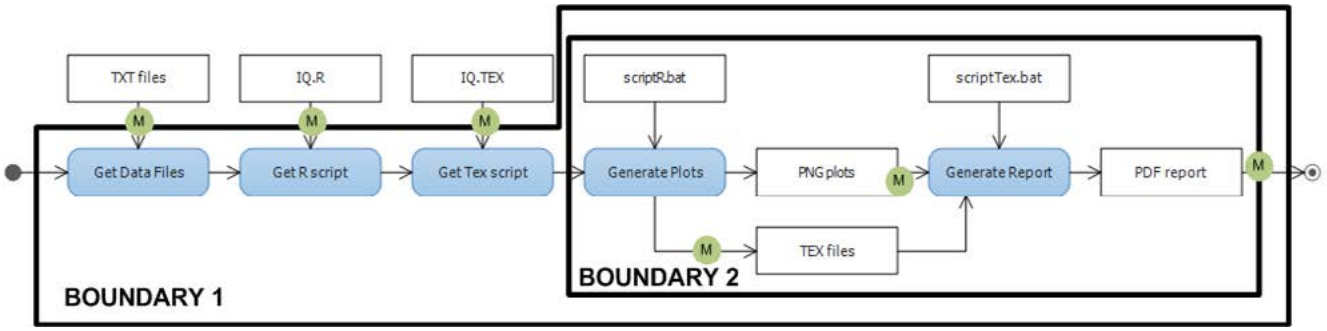


Figure 2: Sequence diagram for eScience process with two process boundaries and measurement points (green circles with 'M') marked.

2.3 Determine internal interactions For both scenarios, the process uses default settings of R and Latex. The scripts are invoking the software from its original locations. However, after careful analysis of software files, it turned out that the default style file of latex *article.cls* was modified. Therefore, this file has to be preserved along with the process and used in the redeployment environment to ensure same effects. Otherwise the final PDF reports will vary in their layout (i.e. number of pages, alignment of content, etc.).

2.4 Ensure deterministic behaviour For the purpose of scenario 1, the interaction between the web service and the process has to be captured and data files provided by the web service have to be provided directly in the redeployment environment. In case of scenario 1 the execution of the process is deterministic and there are no actions needed.

3. Design verification setting In this set of steps, we have identified suitable measurement points enabling us to measure significant properties defined for the redeployment scenarios.

3.1 Specify measurement points The analysis of significant properties resulted in the decision to collect files produced after the execution of each step. The TEX files will be compared with a use of text comparison tools. The PNG plots will be compared in their rendered form. The PDF report will also be compared in its rendered form as well as by examination of its metadata. In case of scenario 1, the communication to the web service will be measured by text comparison of files received from the web service.

3.2 Specify metrics for preservation quality comparison Using goal modelling techniques, we have decomposed high level significant properties into measurable metrics and coupled them with expected target values. For example: the number of pages in PDF report must be equal, the representation of the 'Residuals vs Leverage' in the PNG plot is the same, the duration of calculations is not longer than in the original system, etc.

3.3 Aim for automated measurements capture The analysed process is not formally specified (e.g. by being specified within a workflow engine). All of the process steps must have been performed manually and therefore most of the measurements had to be taken manually. Only in case of the first scenario, the requests to the web service were captured by the tool described in [9].

4. Capture verification data In this step we collected verification data from the original environment by running process instances defined in the first step of the VFramework.

4.1 Prepare system for capturing We have decided to deploy the process in the new clean environment. The system was configured according to descriptions of the process and the required dependencies were added. Thus we validated that all necessary information about the process was collected and there are no interactions which we might have missed.

4.2 Prepare data capture tools In this two steps we introduced tools enabling us to extract verification data from the original system.

4.2.1 Set up tools for capturing external interactions In case of the first scenario, the communication to the web service will be captured by the tool described in [9]. We have also used key logging applications installed in the operating system to collect the inputs from the keyboard.

4.2.2 Set up tools for capturing verification data The data will be collected with the use of tools provided by the operating system.

4.3 Run the process and capture data We have run the instances and collected data from all measurement points in a separate folder. The folder structure and naming convention ensured that the verification data can easily be associated with the executed process instance.

4.4 Verify validity of captured data Due to manual collection of files, each of them was inspected by us before moving to the archive.

5 Prepare system for redeployment We decided to use Ubuntu Linux[1] as a redeployment environment. Ubuntu Linux is an open source project, that is based on the GNU Linux kernel. Ubuntu is easy to use, easy to install, well established and widely used. All of the native operating

[1]www.ubuntu.com

system components are available with open source licenses. Additional packages might be proprietary (such as the Acrobat Reader), but are available free of charge at the moment.

5.1 Prepare system for redeployment We chose Ubuntu Linux 12.10 as the redeployment environment. The operating system was installed with standard configurations within a virtual machine (VM). Our virtualisation environment was VirtualBox[2]. We installed all required updates and ensured the system was up to date. A desktop environment was needed in order to setup the run time environment (see Step 5.1.1).

5.1.1 Set up redeployment System For redeploying the process, we had to analyse which packages are needed in order to substitute the Windows tools that implement the steps of the use case. For the local steps we were able to use the packages available from the Ubuntu repositories. This includes the mathematical statistics package R[3] and Latex[4] (via the texlive package).

5.1.2 Set up external interactions replay to ensure determinism Within our use case there was only one external interaction. The use case needs to retrieve data from a Web service hosted on a machine beyond our influence. In our first redeployment scenario, the Web service was still available and maintained. In the second scenario external dependencies could be removed, as a local data set was used.

5.1.3 Set up data capture tools The tools for capturing the data produced by intermediate steps during the process execution are provided out of the box by the Linux Ubuntu operating system. All intermediate steps produce files that can easily be examined by tools such as diff[5].

5.2 Redeploy preserved process The redeployment step involves executing the original process within its new environment. To achieve successful redeployment, several adjustments have to be performed within the new execution environment. These are described in the following steps.

5.2.1 Identify required preservation actions to enable redeployment

The original client software was implemented with the C#-Language on top of the Microsoft's .NET 4.0 platform, running on a Windows 7 operating system. The .NET platform is exclusively available for Microsoft operating systems. Yet there exist compatible implementations and run time environments for Linux as well. Several attempts have been made in order to redeploy the process within a Linux environment.

First, we tried porting the software client from the Windows .NET 4.0 environment to the Mono Project[6], which is an open source software development platform and run time environment. Mono enables the development and execution of .NET software products, which are binary compatible. Although the mono migration tool[7] indicated full compatibility, direct replacement was not possible due to invalid attempts to access reserved areas of the memory. Having the source code of the client available, we were able to identify incompatible code between mono and .NET[8] implementations within the Web service security stack.

The second approach we considered was porting the client software into a Wine[9] (Wine Is Not an Emulator) environment. Wine allows to run many Windows applications on a Windows platform, by substituting required libraries and acting as a compatibility layer. Hence wine allows to run legacy Windows applications, without the need to maintain the full operating system. Using the package manager winetricks[10] the installation of the required runtime libraries could be scripted. Hence we could use the original Microsoft .Net framework 4 component, that can be installed within the wine environment. This enabled the execution of the client software within the Linux redeployment environment. Hence we could retrieve the data from the web service within a Linux environment.

The next challenge was to orchestrate the packages, that we used for executing the intermediate steps of the process. Adjustments were required regarding naming conventions of applications and paths. Differences occurred in encoding standards and in the scope of included features in the packages. Missing libraries were indicated at the runtime and could be easily installed.

5.2.2 Re-run the set of process instances After the environment has been set up correctly, the use case could be executed. This involved invoking the Web service, which provided the data for further processing. Next, the R script has to be invoked with the retrieved data. The final step produced the PDF document based on the retrieved data.

6 Capture redeployment performance data Data produced during the process execution and captured in the measurement points is collected and verified for its correctness during the course of this step.

6.1 Collect redeployment performance Data The selected measurement points overlap with the outputs of the three steps of the use case which produce intermediate data. This data is persistently stored in files on the hard disk of the redeployment environment.

6.2 Verify validity of captured data The following sub steps aim to verify if the data is not affected by lack of deterministic environment and if the measurements are free of unexpected errors.

6.2.1 Verify if required level of determinism was reached In the first scenario, the Web service is still available. Hence it can be compared to the original data set. In the second scenario, the data is static, hence deterministic.

[2] www.virtualbox.org, version 4.1.26
[3] R, Version 2.13.1
[4] pdfTeX, Version 3.1415926-1.40.10
[5] http://linux.die.net/man/1/diff
[6] www.mono-project.com
[7] www.mono-project.com/MoMA
[8] http:// www.mono-project.com/WCF_Development
[9] www.winehq.org
[10] http://winetricks.org/winetricks

6.2.2 Verify correctness of capture data Once the execution of process instances has finished, the recorded data was verified for its correctness. All of the files could be opened and the screening of their contents was made to verify their correctness.

7. Compare and assess In this set of steps we conducted the comparison of verification data and redeployment performance data. The final assessment about fulfilment of the requirements of the redeployment scenarios was made.

7.1 Compare redeployment performance data and verification data For both scenarios all of the measurement points were available in both environments. We were able to match all of them at the corresponding levels of comparison (see Section 3.1). For example, the TEX files produced by the step Generate Plots were matched for comparison at the file level. Similar matchings were made for other measurement points.

7.2 Conduct preservation quality comparison In this step we calculated the metrics, which were defined in Section 3.2. The examples are: the original PDF report has 169 pages, the new PDF report has 169 pages, values are equal (fulfilled); the process executes in 16,93 s in the original system, the process executes in 12,96 s in the redeployed environment, execution time is not higher (fulfilled), etc.

7.3 Provide summary report Having calculated the metrics, we have created a summary report. It is a document in which all the metrics for each of the scenarios are collected. Clear indication whether the target values are fulfilled is given.

7.4 Make the final decision The report presented that all significant properties of the process were preserved correctly. The requirements of redeployment scenarios were fulfilled. The final decision was made, that the redeployment meets requirements of redeployment scenarios.

7.5 If Positive, remove tools used for verification There was no need to remove the tools.

5. CONCLUSIONS AND FUTURE WORK

In this paper, the VFramework for verification of preserved and redeployed processes was presented. The applicability of the framework was demonstrated on an eScience use case from the domain of sensor data analysis in civil engineering. The preservation and the redeployment of the eScience process was tested by migration to another substantially different environment. For the purpose of redeployment, the process had to be re-engineered and adjusted to work in the new environment. The VFramework was capable of verification of the redeployment in both of the considered redeployment scenarios.

Future work will focus on automation of the verification process. The tools needed for extraction and comparison of measurements taken for significant properties in the measurement points will be created. Furthermore, the VFramework will be tested on further use cases and in different redeployment scenarios.

ACKNOWLEDGMENTS

This research was co-funded by COMET K1, FFG - Austrian Research Promotion Agency and by the European Commission under the IST Programme of the 7th FP for RTD - Project ICT 269940/TIMBUS.

6. REFERENCES

[1] IEEE Std 1012 - 2004 IEEE Standard for Software Verification and Validation, 2005.

[2] ISO/IEC 12207:2008: Systems and software engineering - Software life cycle processes, Feb. 2008.

[3] V. R. Basili, G. Caldiera, and H. D. Rombach. The goal question metric approach. In *Encyclopedia of Software Engineering*. Wiley, 1994.

[4] C. Becker, H. Kulovits, A. Rauber, and H. Hofman. Plato: a service-oriented decision support system for preservation planning. In *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries (JCDL'08)*. ACM, June 2008.

[5] M. Guttenbrunner and A. Rauber. Evaluating an emulation environment: Automation and significant key characteristics. In *Proceedings of the 9th International Conference on Digital Preservation (iPres 2012)*, pages 201–208, Toronto, Canada, October 1-5 2012.

[6] M. Guttenbrunner and A. Rauber. A measurement framework for evaluating emulators for digital preservation. *ACM Transactions on Information Systems (TOIS)*, 30(2), 3 2012.

[7] T. Hey, S. Tansley, and K. Tolle, editors. *The Fourth Paradigm: Data-Intensive Scientific Discovery*. Microsoft Research, 2009.

[8] R. Mayer, A. Rauber, M. A. Neumann, J. Thomson, and G. Antunes. Preserving scientific processes from design to publication. In P. Zaphiris, G. Buchanan, E. Rasmussen, and F. Loizides, editors, *Proceedings of the 16th International Conference on Theory and Practice of Digital Libraries (TPDL 2012)*, volume 7489 of *Lecture Notes in Computer Science*, pages 113–124, Cyprus, September 23–29 2012. Springer.

[9] T. Miksa, R. Mayer, and A. Rauber. Ensuring sustainability of web services dependent processes. *International Journal of Computational Science and Engineering (IJCSE)*, 2013. Accepted for publication.

[10] C. Thanos, S. Manegold, and M. L. Kersten. Big data - introduction to the special theme. *ERCIM News*, 2012(89), 2012.

[11] M. Van der Graaf and L. Waaijers. A Surfboard for Riding the Wave. Towards a four country action programme on research data. A Knowledge Exchange Report, 2011.

[12] Young, R. R. (2004). *The Requirements Engineering Handbook*.

Cloudy Emulation – Efficient and Scalable Emulation-based Services

I. Valizada, K. Rechert, K. Meier, D. Wehrle, D. v. Suchodoletz and L. Sabel,
Albert-Ludwigs University Freiburg
79104 Freiburg i. B., Germany
{isgandar.valizada, klaus.rechert, konrad.meier, dennis.wehrle, dirk.von.suchodoletz, leander.sabel}
@rz.uni-freiburg.de

ABSTRACT

Emulation as a strategy for digital preservation is about to become an accepted technology for memory institutions as a method for coping a large variety of complex digital objects. Hence, the demand for ready-made and especially easy-to-use emulation services will grow. In order to provide user-friendly emulation services a scalable, distributed system model is required to be run on heterogeneous Grid or Cluster infrastructure.

We propose an Emulation-as-a-Service architecture that simplifies access to preserved digital assets allowing end users to interact with the original environments running on different emulators. Ready-made emulation components provide a flexible web service API allowing for development of individual and tailored digital preservation workflows. This paper describes design and implementation of scalable emulation services as part of the bwFLA EaaS framework.

1. INTRODUCTION

Emulation is a key strategy in digital preservation and access to digital artifacts, ensuring that digital objects can be rendered in their native environments and thus maintain their original "look and feel." In most cases the original applications or operating systems developed by the respective software vendors are the best candidates for handling a specific artifact of a certain type [5, 9].

As the number of different past and current computer systems (i.e. hardware architectures) is limited, the number of required emulator-setups is thereby also bounded. Hence, providing access to emulation is suitable for standardized preservation services as well as efficient preservation planning. Nevertheless, deploying full emulation software stacks is a complex and laborious task. Based on these observation, the concept of Emulation-as-a-Service (EaaS) has evolved, aiming towards standardized set of interfaces and uniform access to emulation technology allowing a large, non-technical user-group to make use of emulators and interact with emulated system environments.

Thhis paper's contributions are as follows. We present an EaaS implementation and service model and discuss design issues providing scalable emulation services. We show how *emulation-components* as a core component interact with various emulators and provide necessary APIs and services for data IO like attaching and detaching of virtual removable devices or hard-disks. EaaS users can choose from two different base services: to interact with original environments directly or set up complex preservation workflows. Finally, we present methods for the deployment of EaaS in the cloud (and its scaling on user demand) as well as for user and service authentication in a distributed framework.

2. ARCHITECTURE

The main goal of an EaaS architecture is to develop and maintain a standardized and scalable emulation service model to make emulation a cost-effective digital-preservation strategy and improve its usability. Such a service model, then includes emulated environments either for individual object rendering or represents a component in a larger, complex digital preservation workflow. In contrast to previous projects and approaches to improve usability of emulation technology, the bwFLA project [1] implements a distributed framework. Compared to local provisioning of a complex service stack as proposed in KEEP [3, 2], a networked approach reduces technical and organizational hurdles on the client's side significantly. Instead of adapting a large software package including proprietary software components to various, fast changing end-user devices, the emulators run in a well controlled environment.

The fundamental building block of an EaaS architecture are abstract *emulation-components* (EC) used to standardize deployment and to hide individual system complexity. An EC encapsulates various emulators, available either as open source or commercial products, into an abstract component with a unified set of software interfaces (API). This way, different classes of emulators become also interoperable, e.g. emulators of different vendors could be combined into a larger network compound. The control interface in combination with node- and user-management methods as well as emulator utilities, e.g. for dealing with virtual images, represent a comprehensive EaaS API. Gateway nodes expose the EaaS web-service API. They are responsible for client authorization and authentication as well as for delegating resource requests to the machine management node. The *machine management node* is responsible for efficient hardware utilization, promoting or demoting machines on-demand. To hide complexity of managing dynamic machine allocation, the gateway node also acts as proxy node of an emulation component. The proxy replicates the EC's API, but hides the cloud-specific internal communication from the client. (Fig. 1)

Users need only to implement a single API, which should encourage both interoperability and integration of further, possibly user-contributed, ECs. Furthermore, emulation components are accessible through dedicated web-service (WS)

[1] bwFLA – Functional Long-Term Access, `http://bw-fla.uni-freiburg.de`.

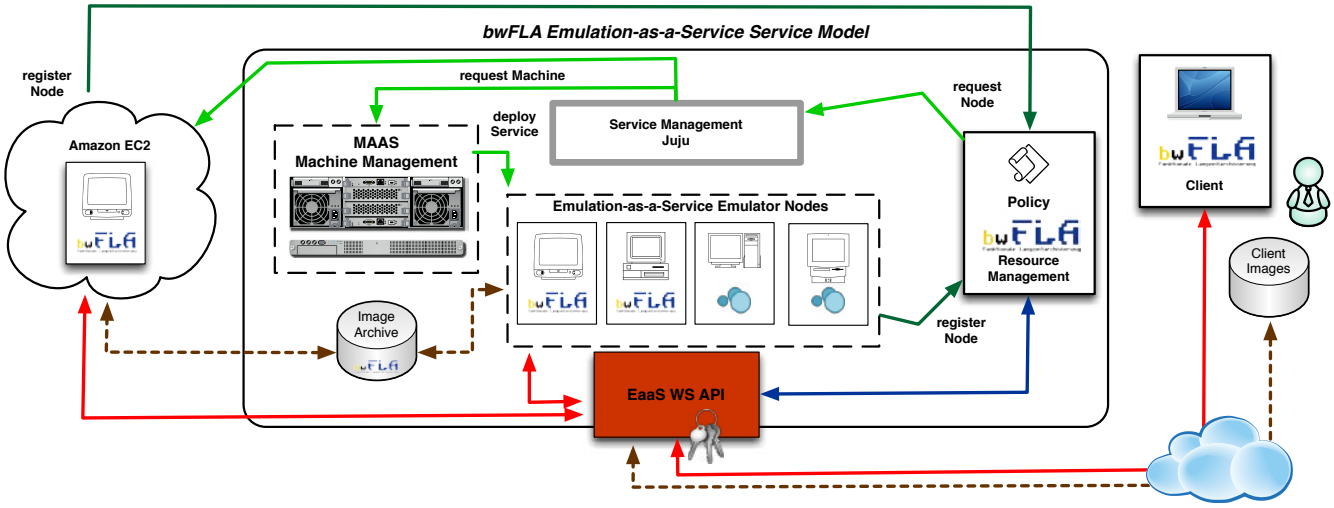


Figure 1: bwFLA: Emulation-as-a-Service General Architecture, Components and Service-Model.

interfaces. This architecture does not enforce specific client implementations. Currently, two variants of clients are available. The user is able to instantiate an EC through a web-front-end and interact with the emulated system interactively. For this option, several standard workflows are already implemented. Such as: bwFLA Ingest/Access/SW-Archive Ingest. The second option is to download the JavaEE-based client framework and build custom workflows.

2.1 Emulation Components Nodes

Emulation-components are implemented as Java EE classes, wrapping a native-platform executable and mapping the emulator's technical capabilities to common interfaces. For instance, every emulator uses a slightly different approach to deal with a set of standard operations like starting and stopping the virtual machine, attaching and detaching virtual drives (floppy, optical or disk drives) or handling network connectivity. Access to the API of any emulation component is possible via its WS front-end. For this a so called WS-service client stub has to be generated via any suitable tool. The generated stub will represent a means of accessing the remote methods of the emulation component, supporting sophisticated client implementations, e.g. in context of specialized workflows.

Currently, the user is able to directly interact with emulated environments using either an HTML5-based web-client or a JAVA-based desktop client. Data-I/O and machine interaction, such as attaching / detaching removable media to the emulation component, is made possible through dedicated utilities in the bwFLA framework and their interfaces.

A future option is to provide dedicated interfaces for non-interactive machine-machine communication, e.g. providing direct access to databases running in an emulated system via network or ODBC interface. A detailed technical description of an EaaS framework and its workflows can be found in earlier work [4, 8].

Finally, to provide a cost-efficient and scalable emulation framework a large scale and especially flexible computing back-end is required. Different emulator types and workloads as well as specific access patterns may require variable computing resources.

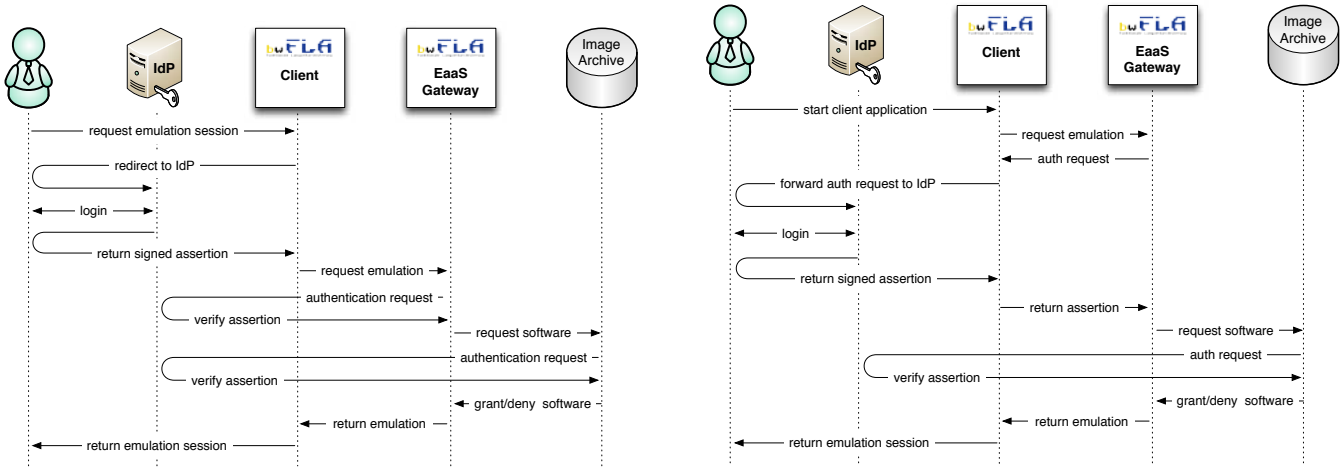
2.2 On-demand Deployment – Scaling EaaS

To become scalable and cost-effective, emulation components need to be deployed only when needed. For this, a suitable framework for hardware- and software-deployment is required. For our purposes, we have chosen Canonical's Metal as a Service (MASS) [2] for hardware management, i.e. for creating emulation component machine instances on demand. If additional hardware resources are required, MASS is responsible for allocation and preparation of suitable machines and installation of a basic operating system (e.g. Ubuntu Linux) on that particular machine. For this, MAAS starts a new physical machine, booting from a DHCP server, downloads and automatically installs the desired OS. Finally, MAAS initializes a user account (e.g. by copying a public ssh-key) for further machine preparation and maintenance. More nodes can be added by just connecting a new machine to power and network. The machine must be capable of booting from the latter using e.g. PXE.

After a machine has been successfully instantiated, in a second step the software deployment system Juju [3] starts installation and configuration of the bwFLA-framework. Juju is an orchestration management tool that requests installed machines from the underlying layer, in our scenario from MAAS. Then it deploys the requested service on that machine by running a (shell) script which installs and configures all needed services automatically. In this way, it is possible to scale a service by requesting additional instances through Juju, e.g. for short-term requirements. If a node is no longer needed, for example, due to a lower load on the cluster, the node is marked as unused and powered down. Hence, the cluster saves energy or the node can be reused for some other services. This service-oriented view abstracts from the underlying hardware and makes the service deployment very simple. A benefit of having the flexible service orchestration tool like Juju is the possibility to use multiple environments for deployed services. Therefore it is possible to have both a local hardware pool managed with MAAS and a commercial

[2] Ubuntu Metal as a Service, `http://maas.ubuntu.com`, version 1.2+bzr1373

[3] Juju, `http://juju.ubuntu.com`, version 0.6.0.1+bzr618



(a) Authentication workflow for web based authentication

(b) Workflow for desktop client based authentication.

Figure 2: Access control in bwFLA Emulation-as-a-Service framework

solution like Amazon EC2.

Since some emulation components require direct hardware access, for instance, CPU virtualization features or CPU ring access, we have chosen a setup which is able to request physical machines as well as virtual ones. In order to reduce hardware costs, commodity hardware is used for the emulation component nodes. The current hardware pool used in out tests consists of standard desktop PCs with a quad-core processor (Intel i5-3470), 8 GByte RAM and a 500 GByte hard disk.

2.3 Provisioning of Legacy Environments

Another important aspect of an emulation service is providing ready-made original system environments, consisting of a basic operating system installation, tailored to be used with a certain emulator type. Typically, emulators provide a set of emulated peripherals and computer components, such as a network chip, sound- and video-card, etc. To make use of these features, appropriate drivers need to be installed and configured. These images act as a base platform, allowing the user to extend them into specialized rendering environments.

Usually, base-images as well as tailored user-images, are kept at specialized institution supplied storage sites thus providing a large variety of systems and specialized software. Furthermore, in some cases, users may choose to not use preconfigured images. In this case, the emulation component should be able to accept a user-provided image directly. For frictionless access through an appropriate emulation component, a suitable network transport protocol is necessary since network quality of service is crucial for usability and performance of the emulated system environment. Especially in cases of user-provided images, network utilization and potentially restricted bandwidth matter. Therefore, a block-oriented protocol has been chosen instead of file-based access since the file representation of a typical emulator image is internally structured as a virtual block-device [7]. A network block device (NBD) and its protocol implement block-layer access over network, i.e. emulating access to physical block-devices, e.g. hard-disk drives. In contrast to file-based access patterns, block-oriented disk-blocks are only requested and transmitted when needed. Thus, an EaaS emulation component becomes immediately operational after initialization. Furthermore, less data transfer might be necessary for instance in case of a sparsely populated virtual disk.

A single virtual disk image may use a few MBytes of storage space for older environments, in some cases up to hundreds of GBytes for newer ones. For efficient and cost-effective maintenance of ready-made images and their user customizations, creating copies for each instance should be avoided. NBD access supports copy-on-write overlays, i.e. allowing for a separation of base-images and user modifications. Any user modification is then stored in a separate block-based differences-file, which can be discarded after session termination or can be kept for future sessions.

2.4 Access Control

To protect resources and, in a second step, support user accounting, a distributed authentication and authorization system is required. Usually, memory institutions already have single-sign-on system deployed, e.g. to protect access to digital publications. These systems were used as starting point for further development. In case of a distributed emulation service, a single central identity provider is not sufficient since users entrust their personal data to their local memory institutions for safekeeping their digital artifacts (e.g. research data). To manage individual user accounts across different sites, a distributed identity management approach is required, delegating authorization and authentication to trusted institutions' identity providers (IdP). Federated identity systems are already successfully integrated at research institutions and universities. For instance, Germany's universities commonly use Secure Assertion Markup Language (SAML) [1] based on identity provider systems, typically Shibboleth [6].

To begin an emulation session a user has to start a client application that is able to communicate to the web-service interface of the emulation component and with the IdP. This can either be a web-site or a program running on the user's computer. The web-site can use the SAML web-login procedure to authenticate the user and access the web-service [10].

The desktop client's ECP module requests an emulation

session from the emulation component on the user's behalf. The emulation component verifies the client's right to use the user's permissions by requesting a signed assertion of the user's IdP from the client. The client's ECP module will forward this request to the IdP and ask the user to authenticate to the IdP (e.g. by providing username and password). The IdP then signs the assertion that grants the user's rights to the client if the user has granted the delegation of his or her rights. The assertion is returned to the emulation component by the client's ECP module. This authorizes the client to interact with the emulation component on the user's behalf.

The emulation component is able to use the same process to request a software image from the archive and return the emulation session to the user if the user has the necessary permissions to access the emulation and the software.

3. RESULTS & DISCUSSION

In order to deploy the bwFLA-framework automatically via Juju/MAAS, some initial effort is required e.g. creating deployment scripts. Installation and software dependencies are to be made explicit and need to be determined upfront. However, as a result, not only a stable and useful service is available but also a documented, reliable and reproducible deployment / installation procedure for other contexts or for future reference created as a by-product.

Currently, emulation components for all major past and present desktop CPU types, PowerPC, Sparc, Motorola 68k, Intel x86, etc., and major operation systems, e.g. OS/2, various MS Windows versions, Apple Macintosh 7.x and newer, etc., have been deployed and can be utilized. Computing nodes running emulation components are either available in cached-mode or need to be created on-demand. If a node is in cache mode it already contains an installed and configured bwFLA framework but is currently inactive. If no more cached nodes are available, 'on-demand' nodes require full installation and configuration. On our available hardware pool, basic node installation and preparation takes about 6-10 minutes plus deployment of the bwFLA framework (2 minutes). Releasing an unused node takes about 1 minute.

4. CONCLUSION

EaaS makes emulation widely available for non-experts. Thus, emulation could prove valueable as a tool in digital preservation workflows, and hence, could become a relevant preservation strategy in many memory institutions. While licensing of past and current software components was not considered in this paper, organizational and technological challenges of emulation as a cost-effective and scalable strategy were analyzed.

The proposed architecture offers both a scalable and easily extendable solution. The scalability of the approach allows the instantiation of emulation nodes on user demand for new emulation resources. The ease of extendability is enforced by the proposed architecture, which is directed towards abstraction of only practically important emulation operations and delegation of their implementation to specific emulator handling classes as well as minimization of effort for adding these handlers in the system's code.

Furthermore, since resources are allocated only on demand, running and maintaining EaaS as a commonly shared infrastructure is efficient in terms of monetary costs, maintenance and management overhead. An unsolved, remaining issue for such a service model is licensing of software. Hopefully, with the availability and the necessity of emulation-based preservation strategies, this issue will vanish.

Acknowledgments

The work presented in this publication is part of the *bwFLA – Functional Long-Term Access* project funded by the federal state of Baden-Württemberg, Germany.

5. REFERENCES

[1] S. Cantor, J. Kemp, R. Philpott, and E. Maler. Assertions and protocols for the oasis security assertion markup language (saml) v2.0. Technical report, OASIS, March 2005.

[2] B. Lohman, B. Kiers, D. Michel, and J. van der Hoeven. Emulation as a business solution: The emulation framework. In *8th International Conference on Preservation of Digital Objects (iPRES2011)*, pages 425–428. National Library Board Singapore and Nanyang Technology University, 2011.

[3] D. Pinchbeck, D. Anderson, J. Delve, G. Alemu, A. Ciuffreda, and A. Lange. Emulation as a strategy for the preservation of games: the keep project. In *DiGRA 2009 – Breaking New Ground: Innovation in Games, Play, Practice and Theory*, 2009.

[4] K. Rechert, I. Valizada, D. von Suchodoletz, and J. Latocha. bwFLA – a functional approach to digital preservation. *PIK – Praxis der Informationsverarbeitung und Kommunikation*, 35(4):259–267, 2012.

[5] J. Rothenberg. Ensuring the longevity of digital information. *Scientific American*, 272(1):42–47, 1995.

[6] T. Scavo, S. Cantor, and N. Dors. Shibboleth architecture: Technical overview. *Working draft*, 1, 2005.

[7] C. Tang. Fvd: a high-performance virtual machine image format for cloud. In *Proceedings of the 2011 USENIX conference on USENIX annual technical conference*, USENIXATC'11, pages 18–24, Berkeley, CA, USA, 2011. USENIX Association.

[8] D. von Suchodoletz, K. Rechert, and I. Valizada. Towards emulation-as-a-service – cloud services for versatile digital object access. *International Journal of Digital Curation*, 8:131–142, 2013.

[9] D. von Suchodoletz, K. Rechert, J. van der Hoeven, and J. Schroder. Seven Steps for Reliable Emulation Strategies – Solved Problems and Open Issues. In A. Rauber, M. Kaiser, R. Guenther, and P. Constantopoulos, editors, *7th International Conference on Preservation of Digital Objects (iPRES2010) September 19 - 24, 2010, Vienna, Austria*, volume 262, pages 373–381. Austrian Computer Society, 2010.

[10] R. Zahoransky, S. Semaan, and K. Rechert. Identity and access management for complex research data workflows. In *6. DFN-Forum 2013 Kommunikationstechnologien*. GI, 2013.

Sustainable Data Preservation using datorium – facilitating a scholarly Ideal of Data Sharing in the Social Sciences

Monika Linne
GESIS - Leibniz Institute for the Social Sciences
Data Archive for the Social Sciences
Unter Sachsenhausen 6-8
50667 Cologne, Germany
+49 221 47694 - 452
monika.linne@gesis.org

ABSTRACT

This paper introduces datorium - a digital data preservation project at the Data Archive of GESIS-Leibniz Institute for the Social Sciences. datorium is a new data repository service for the research community. It functions as a web-based data sharing repository providing a user-friendly tool for researchers making data accessible for the purpose of re-use by other scholars. Sharing, managing, documenting and publishing data, structured metadata and publications will be carried out autonomously by researchers. Data and related information will be available free of charge. All uploaded research data and documentation will be peer-reviewed and digitally preserved by the GESIS Data Archive.

GESIS promotes data sharing as a scholarly ideal and facilitates cooperation between researchers. By developing datorium the Data Archive aims to collect and provide research data with a wide thematic scope for academic re-use. A further intention is to ensure long-term preservation of archived data and metadata as well as providing wide-ranging dissemination possibilities for scholars in order to increase the visibility and availability of their research projects. By providing access to their research data scholars can support new research or secondary analysis and beyond that they profit from increased citations of their work, thereby improving their professional reputation.

Categories and Subject Descriptors

H.3.5 [**Information Storage and Retrieval**]: Online Information Services – *data sharing, web-based service.*

General Terms

Management, Documentation, Standardization.

Keywords

datorium, Social Science Research Data, Research Data Management, Data Repository, Digital Preservation, Data Sharing, Data Archiving.

1. INTRODUCTION

Organizational modification of structural frameworks is demanded thanks to accelerating and fundamental changes within academia and the potential provided by professional information management [1]. By developing a digital data sharing repository GESIS responds to this changing data landscape, where researchers call for flexible ways of distributing and re-using research data. For this reason GESIS is expanding its range of services by offering datorium: a digital data dissemination tool that allows for prompt publishing and sharing of research data with other scholars. In addition, datorium can operate as a working environment that can jointly be used by a research group in order to work together on the documentation of a research project and the publication of related findings.

One goal is expanding the variety of research data types that comes along with a wider thematic collection for data preserved at the Data Archive. With datorium the culture of data sharing, supported and promoted by the Data Archive over the past 50 years, will be pushed forward and facilitated by the re-use of archived data.

A priority of the GESIS Data Archive is to ensure data and metadata provided by the archive is of high quality. Accordingly, all material in datorium is peer-reviewed against defined quality criteria before it can be shared and made available to other scholars.

2. BACKGROUND

Since "*data sharing is essential for all verifications and all secondary analyses*" [2, p.9] producing metadata and sharing research data with other scholars ought to be taken for granted. However, transparency and accessibility of research data is still not common in the social sciences. Archiving and publishing research data in the social sciences is still the exception rather than the rule. This is especially the case for smaller research projects where data is merely a basis for publications not an output in itself. Beyond that, data sets are usually not professionally archived or available to the research community [3, 4, 5].

For this reason datorium focuses on standardized documentation and digital preservation of smaller projects to enhance the

discoverability, accessibility and the reuse of their data, opposed to bigger, national or international survey programs that are well documented, published, and intensively re-used by other scholars. In small research projects the budget often does not cover the cost of archiving and publishing data and metadata. This is a severe problem, as the value of research data is usually not exhausted after the initial research findings have been published [5].

Potential re-use of the data should be considered as part of every research project in order to facilitate further data analysis, such as secondary analysis, reanalysis, replication analysis, verification of research findings [3] or meta-analysis. However, findings of any empirical analysis can only be evaluated if full documentation of the research processes is provided [6]. King (1995) rightly points out, that "*the replication standard holds that sufficient information exists with which to understand, evaluate, and build upon a prior work if a third party can replicate the results without any additional information from the author*" [6, p.444].

In Germany replication and evaluation of research findings from the social sciences is difficult because most publications provide neither data sets nor the syntax used for analysis [7]. This is down to insufficient data management practices in the research routine [8]. Vardigan et al. (2008) specify the problem: *"because good documentation is paramount to effective data use, data archives have long encouraged data producers to document their data thoroughly, starting at the very beginning of a research project and in effect creating an audit trail of all variable transformations that take place over the life of the project. In reality, there is little incentive for data producers to follow these guidelines and documentation is often hastily assembled just before deposit into an archive. Furthermore, documentation is most often produced with word processing software and then rendered into PDF, making reuse difficult"* [9, pp.108-109].

Taking these issues into account, the necessity of developing a data-sharing repository including standardized data documentation is obvious. As Vardigan et al. put it, *"for a secondary analyst to understand a given dataset, he or she must have access to good documentation"* [9, p.108]. Providing this is an essential part of the GESIS Data Archive efforts. datorium will promote this ideal by offering an accessible and user-friendly repository tool.

3. THE PROJECT DATORIUM

Consequently, increasing and intense discussion about Open Access Publication and an attitude shift towards data sharing has emerged. Therefore systems and infrastructures have to be built in order to meet these requirements [1]. In Germany GESIS, in cooperation with other research institutions, have established digital research data initiatives, e.g., the Social Science Open Access Repository (SSOAR), the Social Science Portal SOWIPORT, the social science Literature Information System SOLIS and the social science Research Information System SOFIS.

However, a gap has existed until now for a user-friendly repository which can be managed autonomously by researchers combining upload of data sets with corresponding standardized documentation and metadata. By joining the range of digital research initiatives datorium is closing this gap in the portfolio of GESIS services and the German academic landscape.

As a member of the Leibniz association GESIS is jointly financed by the German federal government and states and it pursues exclusively non-profit objectives. Therefore usage of datorium will be free of charge to data providers and data users. Users of datorium will neither be charged for the upload/download of research data or the review carried out by the Data Archive. This lowers the barrier to unfunded research projects making their data available to the broader academic audience, as well increasing its visibility to an interested public and ensuring its long-term preservation.

Usually access to data in repositories or subject specific data centers is limited to a tight thematic or defined institutional user group. datorium will be thematically open for research data from the wide field of the social sciences and does not restrict the service to institutional members. Thus another barrier can be lowered, allowing researchers to easily publish their research data and related findings.

3.1 The metadata scheme

The metadata schema of datorium is a defined list of structured metadata that gives a standardized description of a research dataset to assist the user. Along with the aforementioned reasons for the necessity of data sharing this ensures an easy way to cite and trace research data. The datorium metadata schema is a simplified version of GESIS's data catalogue (DBK) [10], to which metadata schema from da|ra[1] and DataCite[2] have essentially contributed in the development.

datorium's metadata schema uses mandatory core elements where the user must provide a description of a data set. Additionally the user can choose further optional metadata elements for specification. The metadata schema of datorium is compatible with Data Documentation Initiative (DDI 2) codebook standards and metadata schema from da|ra and DataCite. In addition datorium adopts the metadata standard of the Dublin Core Metadata Initiative[3] that provides core metadata vocabularies in support of interoperable solutions for discovering and managing resources.

By meeting international metadata standards datorium addresses the rising demand for a standardized research data management in the social sciences and serves as a helpful tool for researchers to document their research data in order for it to be understood and re-used by other researchers.

---

[1] da|ra is the registration agency for social science and economic data jointly run by GESIS and the German National Library of Economics, Leibniz Information Centre for Economics (ZBW). This infrastructure lays the foundation for long-term, persistent identification, storage, localization and reliable citation of research data [11].

[2] DataCite is an international consortium founded in London in 2009 comprised of sixteen members from ten different countries, to pursue the common goal of supporting the acceptance of research data as independent citable scientific objects through worldwide uniform standards. On the basis of the DOI-system research data is registered with DOI names to enable comprehensive linking of scientific work with the underlying research data [12].

[3] The Dublin Core Metadata Element Set is a vocabulary of fifteen properties for use in resource description [13]. It is a relevant metadata standard that is commonly used for the description of research data [14].

3.2 Flexible authorization possibilities

Data re-use leads to a decrease in redundant, repeated, data collection and enables more research with more data in less time at less cost.

In certain cases legitimate concerns over data privacy laws, commercial, or national security exist which prevent uncontrolled re-use and prevent data misuse [1]. Therefore datorium offers depositors full control of data and metadata they supply. Depositors have the option to choose a defined category of data access. This means depositors decide who is authorized to access their data. If users wish to access a data set with restricted accessibility, they can request it by clicking an "Apply-Button". An e-mail is automatically sent to the depositor containing information about the person asking for access. The depositor can login to datorium to permit or to deny access to this user. The GESIS Data Archive receives copies of access-requests for the purpose of documentation. However, this also enables the Data Archive to monitor reactions to requests. If the depositor does not provide an answer to a request within 10 to 20 days, the GESIS Data Archive will contact the depositor to investigate possible problems.

Depositors also have the option of data and metadata being preserved only. Here depositors' benefit from the possibility of having data archived by the GESIS Data Archive without providing wider access. If the depositor later decides to publish data, this is easily done.

In detail the access categories are:

a) Free Access: unrestricted download of research data for all registered users, without having to contact data depositors and request access permission.

b) Restricted Access: users have to apply for permission to download the data by contacting the depositor. The depositor manages data access autonomously.

c) No Access: preservation of data and metadata only without publishing. Publishing at a later time is possible.

3.3 Data review

Data and documentation uploaded to datorium is subject to a review process. This is carried out manually, by a 'curator', working at the Data Archive. The review contains technical controls for file formats, data readability, and is checked for viruses. Furthermore, integrity of data and documentation, completeness, data quality, intellectual property and legal aspects are clarified and verified. The curator carries out additional controls for data consistency such as wild codes, missing values, question routing, and weighting factors, etc. A data set is not published until it fulfills review criteria, assuring high quality data is provided.

In the case of rejection, the GESIS curator contacts the depositor and requests correction of the critical content or additional information needed for publishing. Minimum requirements for publishing are specification of the project's title, principle investigator(s), publication year of the data set, and availability status (access category).

As part of the review process all published research datasets and documentation receive a DOI[4] (Digital Object Identifier) in order to:

a) establish easier access to research data on the Internet

b) increase acceptance of research data as legitimate, citable contributions to the scholarly record

c) *support data archiving that permits results to be verified and re-purposed for future study.* [15]

This DOI is published in conjunction with an automatically generated citation that consists of: [primary investigator] ([Year of Current Version]): [Title]. [Data Collector]. GESIS Datenarchiv, Köln. [Study number] Datenfile Version [Number of Version], [DOI].

3.4 Scholarly Collaboration with datorium

Regardless of academic discipline collaborative research is increasingly common. Most social science fields are heading towards cooperative research endeavors. Particularly in sociology the tendency for scholars to work together in the search for systematic knowledge and the understanding of social phenomena is growing [17].

Taking this growing trend into account, multiple users from different locations, regardless of geographic boundaries, can use datorium as a virtual working environment. Researchers are given the option to invite collaborators into a group. This facility helps documentation of research data. Network or project members can communicate through datorium to discuss working progress and track the latest documentation procedures other colleagues of the research group have worked. Subsequent work on documentation can be organized so the burden for each research member is reduced. With datorium interaction and collaboration between researchers is facilitated and supported. Synergies that may emerge from collaboration between scholars who use datorium as a virtual working environment may lead to fruitful social networking options and further research outputs.

4. DATA PRESERVATION

In the first phase of datorium, data and documentation will not be preserved in datorium itself but in the archival storage system of the GESIS Data Archive, where long-term preservation and access to digital objects is provided through file format migration. The Data Archive of GESIS keeps data 'alive' by *"keeping data safe, comprehensible, and secure from physical damage or technological obsolescence so it is available for re-use or repurposing in contemporary or historical research"*[18]. In order to prevent data loss the data archive frequently replaces storage media and checks log files for hardware or software

[4] *A DOI is an acronym for 'digital object identifier', meaning a 'digital identifier of an object'. A DOI name is an identifier (not a location) of an entity on digital networks. It provides a system for persistent and actionable identification and interoperable exchange of managed information on digital networks. A DOI name can be assigned to any entity – physical, digital or abstract – primarily for sharing with an interested user community or managing as intellectual property. The DOI system is designed for interoperability; that is to use, or work with, existing identifier and metadata schemas. DOI names may also be expressed as URLs (URIs)* [16].

errors. Additional scans with checksums or hash functions are carried out to verify bit streams of archived data and documentation remain unchanged [18].

In the second phase, storage of digital objects takes place in the datorium system. Initially, bit stream preservation is used. It is envisioned to later replace this with format migration, depending on the quality and volume of uploaded research data.

5. OBJECTIVES REACHED SO FAR

At the beginning of 2012 the concept and requirement specifications for the repository were generated. The next step was to select a software tool that conformed to the needs of the new repository. After evaluating several open source tools, the decision was made in favor of DSpace[5], since it met most specified datorium requirements.

datorium is presently in its first testing phase (https://datorium.gesis.org, currently registering authorized users only). Interested and authorized researchers are given the opportunity to enter data from their research projects on datorium and autonomously generate metadata. Besides providing initial depositors with an easy way of archiving and distributing their research data, these depositors help develop datorium by providing feedback and suggestions to the GESIS Data Archive.

In order to provide datorium as quickly as possible to the social science community, at present datorium only serves as a tool for the upload of research data, documentation and generating metadata. Currently datorium does not perform as an online platform for publication of research data. For this reason publication of research data is taking place through GESIS's standard distribution system. To ensure high data quality, data and documentation uploaded in the repository must go through the review processes and data preparation procedures carried out by the curator of the Data Archive and described in chapter 3.3.[6] After this immediate review research descriptions are published according to their content through the appropriate retrieval system (e.g., GESIS Data Catalogue, ZACAT-GESIS Online Study Catalogue, Online database HISTAT, Extended Variable Overview, CESSDA Catalogue).

6. WHAT FOLLOWS NEXT

Additional service components will be implemented as further progress is made to the end of 2013. After the rollout of this next stage data producers can upload data sets and publish them through the datorium platform. Here research data will no longer have to be published via the standard retrieval systems of GESIS. For secondary users of data a common retrieval interface will be built that gives an integrated access to both the holdings of datorium and the standard distribution systems of GESIS. At this point the foundation of most of datorium´s targeted aims will have been set – expanding the scope of data types archived at GESIS, to use datorium as a collaborative virtual working environment, and provide user-friendly and fast retrieval via the datorium publication platform.

Beyond this, data producers can apply for data preparation and *added value archiving* carried out by the Data Archive. A*dded value archiving* is set-up for special datasets. For instance, added value data documentation provides extensive (partly multilingual) standardized descriptions of question texts and answer categories, codes and classifications, or interviewer instructions. In addition supplementary contextual information, like comparable questions, codebooks, variable reports or technical reports, is also added. Moreover this elaborate data preparation contains data cleaning, standardization, harmonization, integration/accumulation and enhancement by additional context data.

In the case of added value archiving, storage of research material takes place in the standard GESIS archive and publication is carried out through standard GESIS retrieval systems. This service is also free of charge.

7. ESSENTIAL BENEFITS OF DATORIUM

The benefits of using datorium are various. Some have been mentioned, but it is important to reiterate them, since their advantages should be viewed from different user perspectives:

a) *Benefits for data depositors:*

Data depositors can publish their research data, documentation, and findings free of charge. datorium allows self-funded and small research projects to benefit through increased visibility of their work, with potential citation and an associated enhanced professional reputation. Furthermore, since datorium is a virtual working environment it is possible to conduct collaborative research. Academic partners in multi-partner cooperative projects will be able to produce documentation for research data together in authorized working groups.

Data providers have full control over their data, because by choosing a defined category of data access they autonomously manage access to their data.

Data is published after a quick review to guarantee data and documentation quality. This allows depositors to receive rapid feedback from the research community.

b) *Benefits for data users:*

datorium gives data users free and fast access to the latest research data, which might provide helpful suggestions and inspire new research. Data can be re-used for secondary research, supporting data repurposing. Financially unsupported secondary research can be conducted at low cost, since data collection efforts can be reduced to a minimum. This means data sharing and data documentation permits research findings to be verified and data re-purposed for future research.

As the Data Archive reviews submissions users can be sure they are dealing with high-quality data, without copyright issues or other complicated legal aspects (e.g., observance of data protection).

[5] DSpace is freely available as open source software for academic, non-profit, and commercial organizations building open digital repositories. It preserves and enables easy and open access to all types of digital content including text, images, moving images, mpegs and data sets [19].

[6] This is the reason that metadata covered by datorium follows the metadata schema of the GESIS Data Catalogue DBK [10] and the da|ra Registration Agency for persistent identifiers [11].

c) *Benefits for the academic community:*

datorium facilitates cooperation between scholars and therefore supports synergistic interactions. By publishing data and findings via datorium immediate discussion within the academic community is possible, which might lead to further research based on published data. Uploaded data and related metadata will be preserved long-term, either by format migration or bit stream preservation (depending on the storage location).

Overall datorium has potential to facilitate increased secondary analysis and data re-use.

d) *Benefits for survey respondents:*

Surveys often feel like a burden to respondents. This might be due to reasons of time or for the survey containing sensitive topics that might be hard to deal with. Data re-use eases the pressure on respondents in both cases. Especially data collection from vulnerable groups *"who may be at risk from repeated data gathering intrusions into their lives*" [20, p. 7] can be reduced by data re-use [20]. If researchers re-use data as much as possible they help counteract the effect that respondents become tired or even bothered by taking part in a study.

8. CONCLUSIONS AND OUTLOOK

datorium is suitable for social scientists to efficiently document their research data with associated metadata in a standardized way. Data depositors can digitally preserve their data and share it with the research community. Publishing data and research findings with datorium ensures a visibility to the academic community.

Opening the repository to non-institutional users underlines the originality of datorium's approach, especially so with the focus on small research projects from primary investigators who do not necessarily belong to an institutional organization or are self-funded. A Peer-review carried out by the GESIS Data Archive ensures high data quality. Above this datorium can function as a working environment by allowing multiple partners to jointly create research descriptions within groups.

Because datorium uses a standardized metadata schema interoperable, for instance, with Dublin Core metadata standards it is possible to easily "*weave native Dublin Core Elements into DDI documents*" [20, p.56]. Using DDI enables efficient, accurate use of datasets through standardized documentation. This "*facilitates data access and discovery, improves overall quality, ensures long-term preservation of the information, fosters evidence-based policy making, and supports the establishment of results-based monitoring*" [9, p.108].

By using Dublin Core and DataCite metadata standards datorium meets the conditions for well-organized resource description. By providing datorium to the social science community GESIS promotes the scholarly ideal of data sharing and facilitates long-term digital preservation. Since research data documentation in datorium receives a digital object identifier (DOI), accessibility and traceability of the associated research data is highly reliable.

Implementation is carried out in two phases to provide datorium to the social science community as soon as possible. Since the end of the first phase in April 2013 the GESIS Data Archive provides interested scholars and some of the researchers, who currently have their data documented and digitally preserved by GESIS, access to datorium. In this first phase publication of research data is taking place through GESIS's standard distribution systems, as described in chapter 4. At the end of 2013 datorium will be openly accessible to registered users and research data will be published over the datorium platform (see chapter 5).

9. ACKNOWLEDGMENTS

I would like to thank my colleagues from the Data Archive at GESIS for the inspirational discussions about digital preservation and data sharing. Special thanks go to Wolfgang Zenk-Möltgen, Reiner Mauer, Andias Wira-Alam, Natascha Schumann, Stefan Müller and Laurence Horton for providing their time in numerous meetings as well as their helpful inputs and advices for the realization of datorium.

10. REFERENCES

[1] Winkler-Nees, S. (2012). Stand der Diskussion und Aktivitäten. 2.1 National. In Neuroth, M., Strathmann, S. Oßwald, A., Scheffel, R., Klump, J., Ludwig, J. (Ed.), Langzeitarchivierung von Forschungsdaten. Eine Bestandsaufnahme, pp.23-40. Göttingen: Universitätsverlag.

[2] Fienberg, S.E., Martin E.M. and Straf, M.L. (1985): Sharing Research Data: Report of the Committee on National Statistics. Washington, DC: National Academy Press. [Online: http://www.nap.edu/openbook.php?isbn=030903499X, Accessed 22 April 2013].

[3] Kühne, M. and Meusel, D. (2007): Data Sharing. Unveröffentlichtes Manuskript: Dresden.

[4] Nelson, B. (2009): Data sharing: empty archives. Nature. International weekly Journal of Science 461, pp.160-163. [Online: http://www.nature.com/news/2009/090909/full/461160a.html, Accessed 22 April 2013].

[5] Weichselgartner, E., Günther, A. and Dehnhard, I. (2011). Archivierung von Forschungsdaten. In S. Büttner, H.-C. Hobohm and L. Müller (Hrsg.), Handbuch Forschungsdatenmanagement, pp. 191-202. Bad Honnef: Bock + Herchen Verlag.

[6] King, G. (1995): King, Gary. 1995. Replication, Replication. PS: Political Science and Politics 28: 443–499. [Online: http://gking.harvard.edu/files/abs/replication-abs.shtml, Accessed 1 April 2013].

[7] Schnell, R. (2002): Anmerkungen zur Publikation "Möglichkeiten und Probleme des Einsatzes postalischer Befragungen" von Karl-Heinz Reuband in der KZfSS 2001, 2, S.307-333. Kölner Zeitschrift für Soziologie und Sozialpsychologie 54, 2002, pp.147-157.

[8] Meier, F. (2003): Qualitätsgesichertes Datenmanagement für die Sozialforschung. ZA Information/ Zentralarchiv für Empirische Sozialforschung 52: pp.58-71. [Online: http://nbn-resolving.de/urn:nbn:de:0168-ssoar-199006, Accessed 10 April 2013].

[9] Vardigan, M., Heus, P. and Thomas, W. (2008): Data Documentation Initiative: Toward a Standard for the Social Sciences. The International Journal of Digital Curation. Issue 1, Volume 3, pp.107-113. [Online:

http://ijdc.net/index.php/ijdc/article/view/66/45, Accessed 12 April 2013].

[10] Zenk-Möltgen, W. and Habbel, N. (2012): Der GESIS Datenbestandskatalog und sein Metadatenschema. Version 1.8. GESIS Technical Report 2012-1.

[11] da|ra - registration agency for social science and economic data (2013). [Online: http://www.da-ra.de/en/about-us/, Accessed 15 April 2013].

[12] DataCite (2013). [Online: http://www.da-ra.de/en/about-us/data-cite/, Accessed 12 April 2013].

[13] Dublin Core Metadata Initiative (2013). [Online: http://dublincore.org/metadata-basics/, Accessed 16 April 2013].

[14] Rice, R. (2008): Applying DC to Institutional Data Repositories. Proceedings of the International Conference on Dublin Core and Metadata Applications. p.212. [online]http://dcpapers.dublincore.org/pubs/article/view/945/941 (Accessed 19 January 2013).

[15] http://www.datacite.org/whatisdatacite, Accessed 12 April 2013.

[16] DOI Handbook (2012): The DOI System concept. [Online: http://www.doi.org/doi_handbook/1_Introduction.html#1.6.1 Accessed 18 March 2013].

[17] Babchuk, N., Keith, B. and Peters, G. (1999): Collaboration in Sociology and other Scientific Disciplines: A Comparative Trend Analysis of Scholarship in the Social, Physical and Mathematical Sciences. The American Sociologist 30:5-21.

[18] http://www.gesis.org/en/archive-and-data-management-training-and-information-centre/datenarchivierung/preservation/, Accessed 12 June 2013.

[19] http://www.dspace.org/introducing, Accessed 22 April 2013.

[20] Law, Margaret (2005) “Reduce, Reuse, Recycle: Issues in the Secondary Use of Research Data”. IASSIST Quarterly (Spring). [Online: http://www.iassistdata.org/downloads/iqvol291law.pdf, Accessed 18 June 2013]

[21] Wira-Alam, A., Dimitrov, D. and Zenk-Möltgen, W. (2012): Extending Basic Dublin Core Elements for an Open Research Data Archive. Project Report. In: Proceedings of the International Conference on Dublin Core and Metadata Applications, 2012, S. 56-61 [Online: http://dcpapers.dublincore.org/pubs/article/viewFile/3664/1887, Accessed 18 March 2013].

Modelling Data Value in Digital Preservation

Giuseppa Caruso, Luigi Briguglio
Engineering Ingegneria Informatica S.p.A
via Riccardo Morandi 32
00182 Rome, Italy
{giuseppa.caruso, luigi.briguglio}@eng.it

Brian Matthews
Science and Technology Facilities Council
Rutherford Appleton Laboratory
Didcot, OX11 0QX, UK
brian.matthews@stfc.ac.uk

Calogera Tona,Mirko Albani
ESA-ESRIN
Via Galileo Galilei
00044 Frascati, Rome Italy
{Calogera.Tona, Mirko.Albani}@esa.int

ABSTRACT
For decades, the Earth Science (ES) community has launched missions to monitor vital phenomena of our planet and, through measurements, obtain data for improving their models. Indeed the proper characterisation of phenomena, such as desertification, Arctic sea ice melting, volcanic activities or earthquakes effects, requires the analysis of data acquired in a long period and the validation of correctness of scientific models. This means that digital data, especially in the ES domain, represents an important asset to be preserved over time. Despite each single ES mission's cost being quantified and supported by well documented evidence, ES organisations are not able to assess the value of data generated by those missions over time. This paper describes the rationale for and an approach to modelling the value of data/information to be preserved over long term in digital archive. This is the result of experience in the SCIDIP-ES project [16] which has considered the: i) definition of models for describing the value of digital data and related information; ii) characterisation of data/information value model through core set of key parameters and iii) identification of long term digital preservation activities that may potentially impact on key parameters and consequently on the value of digital assets. This model is being assessed in ES scenarios with data curators and archive managers.

Categories and Subject Descriptors
H.1.1 [**Information Systems**]: Systems and Information Theory – *value of information.*

General Terms
Management, Economics, Theory

Keywords
Value of Data/Information, Value Model, Sustainability, Long Term Data Preservation, Earth Science (ES).

1. INTRODUCTION
Climate change is arguably the greatest environmental challenge facing us in the twenty-first century, and this has been recognized in reports from the Intergovernmental Panel on Climate Change (IPCC) [19] and from the United Nations Framework Convention on Climate Change (UNFCCC) [20]. The consequences of a warming climate are far-reaching, potentially affecting fresh water resources, global food production and sea level. Threatening impacts on the natural environment and life on Earth for generations to come, climate change is high on political, strategic and economic agendas worldwide. This premise highlights the importance of ES studies and describes the Earth and its natural phenomena through data and models. For this purpose, ES community - which includes a wide range of scientists interested on fields related to the Earth such as physical geography, geology, meteorology, oceanography, atmospheric sciences, physics, and chemistry - acquires, processes and examines a large amount of dataset on Earth's materials, structure, history and all of the living things on it, including how and when they formed and evolved. This kind of study of the Earth helps to develop an understanding of its future and the need for careful management of its resources, and in particular, this can help to model and estimate climate change. For those reasons, for decades ES community launched missions such as Argo [21] and GRACE [22] which acquire data related to gravimetry and Mean Sea Level variations, very sensitive indexes of climate change and variability. It is also to be considered the large amount of new ES observations upcoming in the next years will lead to a major increase of ES data volumes, as well as ES datasets are characterised by heterogeneity due to different instruments and technologies mounted by each mission's satellite. It is important to highlight that validation and improvement of models cannot be successfully performed in case of "lack" or "hole" within the dataset sequence. In other words, every acquired data from the ES missions is an important asset for ES community and the whole humanity: that clarifies the importance of avoiding the loss of data related to Earth events uniquely occurred over time and space, as well as to plan and enact long term digital preservation on this asset for ensuring availability and accessibility.

An asset for an organization has, for definition, a value. While costs for generating data are widely known and documented, on the other hand, it is still an open issue for ES organizations to assess the value over time of this asset. This paper describes in Chapter 2 the existing models available from the state of art and their limits in satisfying specific needs of the ES community, especially when dealing with long-term preservation. In order to overcome those limitations, Chapter 3 introduces the experience carried out within the SCIDIP-ES project which provides an approach for adapting existing models and describes how those models have been extended. Closing remarks are reported in Chapter 4.

2. VALUE OF DATA/INFORMATION

The term, "Value" has multiple meanings, which change according to the different domains (Sociology, Economics, Ethics [23]) where this term is used. In this paper, the term "value" is for referring to the economic and market value of preserved information, which is seen as an asset. In economic studies, the theory of value attempts to explain the exchange "value" or "price" of goods and services.[27][29][30]. According the Marketing approach, the "value" may be conceptualized as the relationship between the consumer's perceived benefits and the perceived costs for receiving these benefits[24][25][26]. From the point of view of the profit and no-profit organisations, the generation of value depends on the difference between benefits and costs derived from their activities[31][28].

The value approach followed in the SCIDIP-ES project and presented in this paper is near the last one, considering that the *value of data and in particular of the preserved data is closely related to processes and activities, which are needed over time to offer the data/information to final users as well as to the activities performed on data/information by data users*. In this perspective, Benefit/Cost analysis is the starting point for the value analysis and how it changes during the whole digital object lifecycle. Thus, to achieve a better understanding of relevant current and past work on benefit/cost analysis and on the Value Analysis about information and in particular preserved information was an important step to identify existing value approaches, which could be followed by the SCIDIP-ES project.

2.1 State of the Art

The interest in digital preservation and its value is evident through the relevant related work. However, the most of the analysed research projects on digital preservation have been focused on the Cost Model and on, in particular, the estimation of their cost. Those analyses have been carried out in different domains, with a particular focus on culture heritage. It is characteristic that cost models for digital preservation take a ***lifecycle approach*** (LIFE [1], CMDP [2], KRDS [3] , ENSURE [4]). However, no common consensus has yet been reached on how the lifecycle for costing digital preservation should be structured; or on how the individual lifecycle phases should be broken down and detailed, perhaps due the high dependences of preservation costs on the range of services that an institution can offers. All the considered projects adopt the OAIS reference model [5] as starting point for the definition of digital preservation lifecycle and its breakdown but the final results of the latter are quite different among those projects, due to the different fields of application. Another unresolved or hidden issue is the development of formulas for operational cost models.

With regards the Value analysis, the studies [6][7][9] dealt with about the general value of the Information for society; they are not about the preserved data but more in general on the impact of it on the domain where it is used. However*, all are persuaded that* ***the value*** *of information* ***depend on its use and its capability to be shared***. Keeping Research Data Safe[3] (KRDS) is the only study to consider the benefit analysis for data preservation, which also provides a Benefits Analysis Toolkit [8]. This latter has been tested, reviewed and developed further in the Keeping Research Data Safe (KRDS) Benefits Framework and the KRDS/I2S2 Value Chain and Benefit Impact Analysis tools for assessing the benefits of digital curation/preservation of research data. In conclusion, from this analysis of related work, a list of variables and parameters was defined. This paper does not include that list which is available in the project's document [11] , but it is relevant to highlight at least the typologies of variables/parameters identified. In fact, two main typologies of parameters were identified: those related to cost analysis and its definition (in this perspective it is possible to define the value of the preserved object as sum of the cost elements); the others one are general and high level parameters about digital object quality and features.

Finally, the main identified Economic Value model approaches relevant for value analysis of the Data/Information, were:

1. *Willingness Approach*: the Value of Information(VoI) measured according the willingness to pay of decision-makers (or others who use the data) where their willingness depends on the level of uncertainty and on what is at stake (amount of possible loss without information)
2. *Attribute Approach*: the value is a function of some parameters related the quality and features of the digital bject;

   VoI *= f (Usability, Shareability, Time, Accuracy, Precision, Risk, Unicity, Integrity)*
3. *Historical Cost Approach*: VoI as approximation of the cost of acquiring/creating/archiving/preserving it (purchase price or development cost);
4. *Present Value Approach*: information considered as an asset is valued based on the present value of expected future economic benefits.

The first two are market oriented, that means that they define the value according the value perception of who use the product or services according their features and quality, and the user's availability to pay or to do something in order to access to the asset. The last two approaches are process oriented, that means that the value of the provided product or service is defined according the process and cost for providing it as well as the produced benefits in terms of outcomes derived from an activity or work process.

2.2 Limits of current models

The state of the art analysis gave an overview of the available and more used approaches about the value analysis as well as provides for the SCIDIP-ES project an early idea about their advantages and limits according the project needed. In this perspective, it is possible to highlight the following aspects:

- Most of the value models analysed may not be applied to Preserved data, because they are mainly focused on cost analysis.
- Those models are not addressing the benefit provided by the data itself, that is considered an important aspect for the ES community and consequently for the SCIDIP-ES purpose.
- Moreover, current experiences are not considering the whole lifecycle of digital data which may impact on its value.

Starting from the models identified, it becomes important to adapt and extend them, for the specific purpose of the project. In order to achieve this goal, the SCIDIP-ES team proposes to adapt and extend:

- The Historical Cost approach by adopting for the cost analysis ABC (Activity Based Costing) model and introducing a benefit framework for the benefit analysis;
- The Attribute Approach by introducing the SCIDIP-ES core set of preservation parameters, which allow the definition of the value of data/information and the impact on this value due to activities performed on data during the whole lifecycle.

3. VALUE OF PRESERVABLE DATA

The proposed model aims to bring together both to the process oriented approaches and to the market one. In this perspective, this section offers more details about the Cost/benefit framework as a process oriented approach as well as on the extension of the attribute approach as a market oriented approach.

3.1 Tailoring Benefit/Cost Analysis

Cost-benefit analysis takes into account the positive and negative aspects related to a case to be evaluated. Those aspects must be expressed in terms of a common unit of value, which conventionally is money. That represents a limit for measuring the benefits generated from the long term digital preservation activities in the scientific domain, since currently most of the data and information are freely available for users. Thus the benefit analysis proposed in this paper suggests measuring them following an approach based on the identification of the general impact on the community and society. With regards the cost analysis based on the Activity based Costing Model, the main effort, to tailor it, was to define an activities Framework for digital preservation relevant for scientific organisations.

3.1.1 Analysing Data Benefits

The following section will go deeper into the benefits of the data product. This approach starts from the analysis of the KRDS [3] benefits model, before passing to a more systematic model to be applied to data product relevant to scientific data.

The KRDS model of benefits [8] defines 3 dimensions: outcomes, timescales and beneficiaries as a framework to evaluate the benefit of a data product. Outcomes are then divided into:

- Direct benefits: positive impacts obtained in a data curation activity.
- Indirect benefits: negative impact avoided by investing in a data curation activity.

The guide to the benefits framework then goes on to discuss how this framework might apply in particular instances. This gives particular instances of outcomes which might apply; however, these are an unstructured list of potential outcomes.

In the SCIDIP-ES project a more systematic characterisation of the outcomes is proposed which could be applied to a data product within a research data scenario. This approach can then be combined with the rest of the KRDS approach to provide a more detailed analysis of the potential benefits accruing from the preservation of a data product.

This approach can also be compared with that of Whyte and Wilson [14] who identifies seven general criteria for retention (*Relevance to Mission; Scientific or Historical Value; Uniqueness; Potential for Redistribution; Non-Replicability; Economic Case; Full Documentation*). Again, while these are useful, they are not comprehensive, and do not in general capture the intentionality behind the criteria which may lead data archivist to identify additional benefits not covered within these definitions, or provide measurable criteria.

The nature of the benefits can be analysed by considering two main categories of benefits: *Utility* and *Substitutability*. These categories approximately correspond to KRDS's direct and indirect benefits.

Substitutability factors are those which assess whether an alternative data set of an acceptable quality which can be used in place of the data can be accessed if it is needed, if the archive's copy is not available. If a reasonable substitute can be accessed elsewhere, or generated afresh at a reasonable cost (for example at a lower cost than continuing to preserve the data), then the benefit of keeping a copy of the data within the archive is likely to be lower.

Utility factors consider the value of the data for re-examination and reuse in the future. Thus if the Utility of the data is high, then the benefit of the data is high. Considering data utility further, clearly the data is more valuable if the data is *desirable,* that is it requested, re-examined and reused in the future, especially in new contexts and new situations. Data may also have more beneficial impact if it is *reusable*, that is presented in a manner which encourages re-examination and reuse; if it is easier to comprehend and to integrate with other data and computing systems, it is likely to be reused, and thus have a higher utility. To this end, some instances of the types of evidence for the benefits of data in terms of both substitutability and utility have been identified, together with some guidelines on metrics which might be used by a data archive to measure such evidence. Those evidences and metrics bring together the concepts to estimate in terms of benefits, the gross value of the data. It is important, anyhow to highlight that often such metrics are subjective and difficult to measure, especially for a long time in the future. For brevity, we omit a comprehensive treatment here; Table 1 gives some examples of evidence of Data Desirability.

Table 1. Data Desirability Metrics

Evidence	Description	Metric
Data requests	Number of requests for the data arising from the user community.	Number of user requests. This can be also measured by a percentage of the funding which is supporting the user community (e.g. future research grants).
Data Citations	Citations of the data within refereed published literature.	Number of citations to data (or a reference paper for the data), weighted by the impact factors of the citing papers.
Research grants	Future research grants which cite or request access to the data. This is evidence that the data remains relevant in an active research area.	Percentage of the value of research grant.
Commercial data access	Sales of access to the data or added value products using the data.	Value of sales of the data or derived products.
Patents	Use of the data leads to commercial patents.	Number of patents arising (and an estimate of their value e.g. use in products).
Products	Use of the data leads to commercial patents.	Value of sales of products.
Influencing decisions makers	Use of the data by government or other agency to either: - influence policy (e.g. included in IPCC report) - directly influence action (e.g. monitoring of volcanic ash and flights)	Citation of data in policy documents. Estimate of value of policy or action.

3.1.2 Analysing Data Costs

Estimating the cost for long-term digital preservation has received attention from many organisations (e.g. companies, digital libraries, research data centres) who are interested in preserving for their data. In Earth Science domain, this interest is due in particular to some data attributes as the non replicability of the acquisition process within the same conditions (i.e. satellite or airborne data), which could lead to the loss of relevant data as well as to the loss of the cost for generating them, in absence of an appropriate digital preservation strategy. In addition, this interest is because a sound cost model should lead industries to better understand economic impact of digital preservation. Despite that, cost modelling for long-term digital preservation is a relatively new area of study. Many research projects analysed above (e.g. Life Cycle Information for E-Literature (LIFE)[1], Keeping Research Data Safe (KRDS)[3] and NASA's Cost Estimation Tool (CET) [15]), dealt with the cost model. Those existing studies are related to specific projects, institutions or materials and therefore difficult to transfer into other contexts. That is due to the particularity of the costs of preservation which are determined for specific digital assets using specific technologies, at a specified level of reliability and so on. From that perspective, it may be possible to follow the approach and high level model of others experiences, while tailoring them according the specific case requirements.

For the SCIDIP-ES project's needs for costs analysis it has been decided to follow an approach **based on Activities Based Costing (ABC) model**, which seems the most frequently used approach for the cost analysis. This is a costing methodology that identifies activities in an organization and assigns the cost of each activity with resources of all products and services according to the actual consumption by each activity. In that perspective, it is powerful tool both for cost assessment and for better understanding organisation processes. For such reason, this method is very useful to: i) identify and eliminate or modify production or service processes that are ineffective; ii) support an economic analysis of the adoption of new production or service processes. The first step in designing an ABC system is to conduct an activity analysis to identify the resource costs and activities of the organisation. The activity analysis identifies the work performed by the organisation to carry out its operations. Consequently, activity analysis includes gathering data from existing documents and records, as well as collecting additional data using questionnaires, observations, or interviews of key personnel. In our specific experience, we have identified the activities through two ways:

- In the first part of the analysis, the high level activities have been defined according the past experience of other projects which provided their cost models and some approaches for the breakdown of the activities for organisation committed in the digital preservation (e.g.: LIFE, KRDS, ENSURE);
- Then a re-adjustment and an identification of other lower level activities more related to Science domains has been carried out through internal discussion and analysing in particular the current digital preservation process inside ESA (European Space Agency).

For each high level activities group, two other levels of sub-activities were defined. The activities classes and groups are significant for economic assessment of the different parts of the overall system which brings a product or a service to a customer. This activities model represents the most important part in the ABC model application. The high level activities are conceptually based on the OAIS reference model [5] following the approach of many other projects (e.g. [1][2][4][17][18]) engaged in the cost analysis. The lower levels are more related to science organizations; they form a guide for users and can be contextualised to the structure and language of the organisation.

The Figure 1 shows the first and second level of the identified activities proposed for our cost analysis scope.

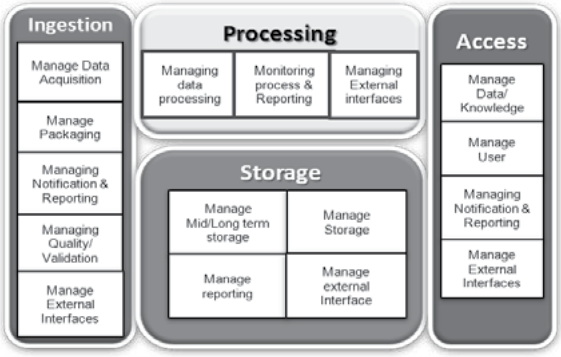


Figure 1. LTDP Activities 1st and 2nd levels –ABC approach

3.2 Extending Attribute Approach: Core Parameters

From the analysis of the parameters identified in the state of the art as well as from discussion within the SCIDIP-ES project, is come to light that some data's features are very relevant to explain the engendering of value of data itself. The selected features which are by us called "core parameters" are a core set of five parameters, which qualifies the value of data/information according the Data Users. They are defined as "factors that characterise the preserved digital object, which could impact on the utility perception of who needs and uses the digital object". Consequently they influence decisions on data use by data users impacting on the benefits generation.

On the other hand, providing over time digital data with the required degree of core parameters means to be aware on the organisational activities and resources (e.g.: technologies, know-how), which impacting on them as well as to be able to leverage on activities and resources for achieve the required levels of those parameters.

Those parameters are defined as follow:

1. Availability

Availability is the property that a data is available for long-term use and at the time it needs to be utilised.

Data availability (sometime related to the concept of timeliness [12]) is one of the most frequent data quality dimensions that must be managed. According Vermaaten, Lavoie and Caplan [13], in order to ensure availability, the digital object must be ingested into, and subsequently maintained by, a preservation repository.

2. Accessibility

Accessibility is the ability to access data from some system and/or entity. Accessibility requires rights and/or permissions to access the data, technology (i.e. hardware and software) to access the data and the related documentation necessary to understand the data itself. In some case the data could be available but its access is not possible or not easy. This reduces its value for the interested community because becomes difficult to use it.

3. Integrity

Data Integrity is defined as the ability to ensure that data is not altered or destroyed in an unauthorized manner. This complies to the ISO:14721:2003 OAIS definition [5].
Usually we could say that enforcing data integrity ensures the quality of the data. Data integrity refers to maintaining and assuring the accuracy and consistency of data over its entire life-cycle. The data integrity is very important in particular in the business, administrative and legal domains as well as in science and research because this feature assures the reliability and trustworthiness of result derived from data itself.
Data integrity imposes a strong commitment on the organisation involved in the data curation and preservation, by adopting well defined rules of actors involved in the processes, as well as standards and procedures. But to provide data assuring its integrity allows improving the utility for the Data users and consequently the benefits.

4. Completeness

Data completeness is defined as the degree of data to be provided with all the comprehensive and correct information in order to facilitate future discovery, access, and reuse. That includes any description on the resource's provenance and the context of its creation and use. This is a data quality dimension dealing with how complete the data is. In any data resource, it is essential to meet requirements of current as well as future demand for information. Data completeness assures that the above criterion is fulfilled.

5. Usability

ISO 9241 defines usability as "The extent to which a product can be used by specified users to achieve specified goals with effectiveness, efficiency, and satisfaction in a specified context of use."
The usability improves the capability to compare, correlate and aggregate set of data. Usually usability of set of data is assured by the adoption of common standard and methods. In terms of process cost, of course providing usable data means to have defined preservation plan, standard and method agreed with community.

3.3 Preservable Data Value Model

The SCIDIP-ES Value model (fig. 2 and 3) in order to overtake the mentioned limits of the other models (par. 2.2), has tailored the benefit/cost analysis, extending it with the adoption of the attribute approach. The inclusion of them in that model is important since they identify the quality level required for guaranteeing the usage of the data over time, at which are closely related the generation of benefits as well as of the organizational costs. The former is performed by the proposed benefit framework as well as by the data activities analysis for the cost analysis based on the ABC model

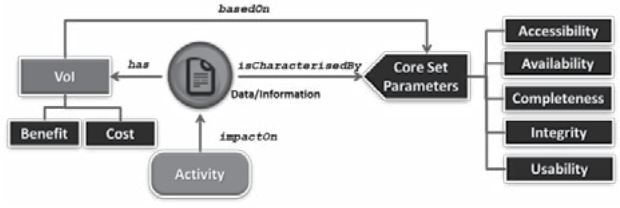


Figure 2. Value of Preservable Information Model

In this perspective, that model also takes in consideration the main relevant users: data provider/manager and data user. For this reason, the model has been extended by considering the specialization of activities carried out/controlled by two users, as shown in Figure 3.

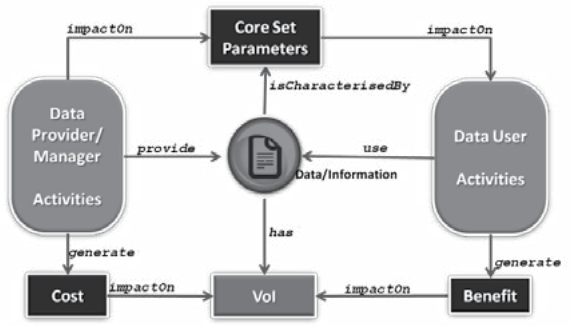


Figure 3. SCIPID-ES Value Model

This model is centred on the data and starts from the consideration that data has a value (VoI), which is determined and impacted from benefits and costs. The benefits are generated by usage of data, while the costs are generated from the activities performed from the data provider/managers in order to maintain and provide the data itself. Moreover, the data is characterised by attributes (so-called parameters), which impact on the utility perception of the data user. Indeed, that latter will decide to use data according to values assumed by core set parameters and acceptance thresholds/criteria. It is important to highlight that acceptance thresholds/criteria may differ between different organizations, based on their internal policies and objectives. However, improving those attributes, according this model, means to increase the probability that data /information will be used over the time, increasing consequently the possibility to generate more benefits. On the other hand, the data provider/manager activities impact on values assumed by core parameters for each data set provided. Consequently, data provider/managers should keep in mind those core parameters when plan or perform activities and choose resources (e.g.: technologies) for preserve digital data.

4. CONCLUSION

The paper addresses the issue of assessing the value of digital asset for ES community, that is the huge amount of data available from a variety of ES missions and preserved in ES archives. Of course, this is a crucial point also for other fields as Social Science, Bioinformatics, Astronomy, Particle Physics, Medicine and Health, where the quantity of information that will be stored in digital form will increase dramatically.

This amount of data has to be preserved and the most difficult task to be performed by data owners is the assessment of its value. It cannot be derived from just the cost of missions, because that is a component which takes into account the only generation aspect, while beyond data generation it has to be considered the whole lifecycle and performed activities on data itself. On this perspective, this paper has described the existing models from the state of the art for assessing the value and those models have been analysed for identifying limitations in supporting data owners. Consequently, in order to overtake those limits, it has been described the proposed approach for adapting the existing models, mainly based on historical cost approach (process oriented). Moreover, it has been enriched by including the benefit framework and by analysing the contextualised activities for cost definition, according to ABC model. Finally, the model has been extended by characterising the data through a core set of parameters which may potentially impact on value of data itself.

This model is being assessed in ES scenarios with data curators and archive managers, in order to carry out an economic sustainability analysis of: i) the Long Term Data Preservation (LTDP) in the ES domain as well as, ii) the developed SCIDIP-ES Infrastructure which provide a set of services and toolkits for managing digital preservation of ES-data.

5. ACKNOWLEDGMENTS

Work partially supported by the European Community under the Information Society Technologies (IST) program of the 7th FP for RTD - project SCIDIP-ES, ref. 283401.

6. REFERENCES

[1] LIFE project - Life Cycle Information for E-Literature (2005-2010) http://www.life.ac.uk/

[2] CMDP Project - Cost Model Digital Preservation . 2010-2012. http://www.costmodelfordigitalpreservation.dk/

[3] KRDS Project - Keeping Research Data Safe. 2007- 2010. http://www.beagrie.com/krds.php

[4] Ensure Project - Enabling kNowledge Sustainability Usability and Recovery for Economic value .2011-2014. http://ensure-fp7-plone.fe.up.pt/site/

[5] The Consultive Commitee for Space Data Systems: Reference Model For An Open Archival Information System (OAIS), CCSDS 650.0-M-2, 2012. http://public.ccsds.org/publications/archive/650x0m2.pdf

[6] Molly K. Macauley. The Value of Information: A Background Paper on Measuring the Contribution of Space-Derived Earth Science Data to National Resource Management. Discussion Paper 05–26. May 2005. http://rff.org/rff/Documents/RFF-DP-05-26.pdf

[7] D. Moody, P. Walsh. Measuring The Value Of Information: An Asset Valuation Approach. European Conference on Information Systems (ECIS'99). 1999.

[8] Charles Beagrie Ltd. Guide to the KRDS Benefits Framework. Keeping Research Data Safe. report, v3. July 2011. http://www.beagrie.com/KRDS_BenefitsFramework_Guidev3_July%202011.pdf

[9] S. Schwolow , M. Jungfalk . The Information Value Chain Strategic Information Management for Competitive Advantage . Bachelor's Project at Copenhagen Business School. 2009. http://www.informationvaluechain.com/wp-content/uploads/information-value-chain.pdf

[10] Mark Bide, Mark Bide & Associates - Business Models For Distribution, Archiving And Use Of Electronic Information: Towards A Value Chain Perspective - A Study For ECUP+ . 1999. ftp://ftp.cordis.europa.eu/pub/ist/docs/digicult/businessmodels.pdf

[11] G. Caruso, B. Matthews, B. Polsinelli . Parameters for Long Term Preservation and data sustainability model. SCIDIP-ES-DEL-WP31-D31.1. April 2013

[12] Heinrich, Bernd, Klier, Mathias. A Novel Data Quality Metric For Timeliness Considering Supplemental Data. 17th European Conference on Information Systems 2009. http://www.ecis2009.it/papers/ecis2009-0656.pdf

[13] Sally Vermaaten, Brian Lavoie and Priscilla Caplan. Identifying Threats to Successful Digital Preservation: the SPOT Model for Risk Assessment. September/October 2012 D-Lib Magazine http://www.dlib.org/dlib/september12/vermaaten/09vermaaten.html

[14] Angus Whyte and Andrew Wilson. Appraise & Select Research Data for Curation. Digital Curation Centre and Australian National Data Service "working level" guide, 25 October 2010. http://www.dcc.ac.uk/resources/how-guides/appraise-select-data

[15] NASA's Cost Estimation Tool - http://opensource.gsfc.nasa.gov/projects/CET/index.php

[16] SCIDIP-ES project – SCIence Data Infrastructure for Preservation with focus on Earth Science. 2011-2014. http://www.scidip-es.eu

[17] Anna S. Palaiologk, Anastasios A. Economides, Heiko D. Tjalsma, Laurents B. Sesink : An activity-based costing model for long-term preservation and dissemination of digital research data: the case of DANS , International Journal on Digital Libraries, Volume 12, Issue 4, pp 195-214, September 2012. http://link.springer.com/article/10.1007%2Fs00799-012-0092-1

[18] Beagrie, N., Chruszcz, J., & Lavoie, B. : Keeping research data safe: A cost model and guidance for UK universities. Copyright HEFCE 2008. http://www.jisc.ac.uk/media/documents/publications/keepingresearchdatasafe0408.pdf.

[19] IPCC - Intergovernmental Panel on Climate Change. http://www.ipcc.ch/

[20] UNFCCC - United Nations Framework Convention on Climate Change. http://unfccc.int/

[21] ARGO project . 2000-on-going http://www.argo.net/

[22] GRACE – Gravity Recovery and Climate Experiment, since 2002. http://www.csr.utexas.edu/grace/

[23] George Macesich, Dan Voich, Lee P. Stepina and Dan Voich Jr. Cross-Cultural Analysis of Values and Political Economy Issues, 1994

[24] Peter Doyle. Value-Based Marketing: Marketing Strategies for Corporate Growth and Shareholder Value. Wiley, 2000

[25] Raquel Sánchez-Fernández, M. Ángeles Iniesta-Bonillo. "The concept of perceived value: a systematic review of the research," Marketing Theory 7, 2007

[26] Philip Kotler, Gary M. Armstrong. Principles of Marketing, Prentice Hall, 2010

[27] Howard Nicholas, Marx's theory of price and its modern rivals. London, Palgrave Macmillan, 2011

[28] R.Brealey,S.C.Myers,F.Allen. Principles of Corporate Finance, McGraw-Hill/Irwin 2010

[29] Alexander Gersch. On the Theory of Exchange Value. 1969

[30] Steve Keen. Use-Value, Exchange Value, and the Demise of Marx's Labor Theory of Value. Journal of the History of Economic Thought / Volume 15 / Issue 01 / Spring 1993, pp 107-121

[31] Harry F. Campbell,R. Brown. Benefit-Cost Analysis: Financial and Economic Appraisal using Spreadsheets [Paperback], 2003

The process of building of a national trusted digital repository: A user centric approach for requirements gathering and policy development

Aileen O'Carroll
Digital Repository of Ireland (DRI)
National University of Ireland, Maynooth
Co. Kildare – Ireland
aileen.ocarroll@nuim.ie

Sharon Webb
Digital Repository of Ireland (DRI)
National University of Ireland, Maynooth
Co. Kildare – Ireland
Sharon.webb@nuim.ie

ABSTRACT
In this paper we describe a process of consultation and data gathering with key stakeholders conduced by the Digital Repository of Ireland (DRI) in 2011/2013. This paper will examine the contributions of the interview process to policy development and requirements gathering with a particular focus on access, reuse and community engagement.

Keywords
Case Studies and Best Practices: Processes, Metadata, Systems, Infrastructure, Community, Policy, Requirements.

1. INTRODUCTION
The Digital Repository of Ireland is an interactive national trusted digital repository for contemporary and historical, social and cultural data held by Irish institutions; providing a central internet access point and interactive multimedia tools, for use by the public, students and scholars. It is a four-year exchequer funded project, comprising six Irish academic partners, and is supported by the National Library of Ireland, the National Archives of Ireland (NAI) and the Irish national broadcaster RTÉ. A key task is to link together and preserve the rich data held by Irish institutions, provide a central internet access point and interactive multimedia tools. Enabling access and reuse to research data is a central challenge. This article outlines how the process of qualitative interviews conducted by DRI allowed us to develop a complex understanding of the barriers which might limit the ability of data to be shared. An unexpected outcome of this process was that it facilitated community engagement. This assisted in developing the relations of trust which are so important for overcoming barriers to access and data sharing.

2. METHODOLOGY
In this section we outline the contribution of qualitative interviews to requirements gathering and policy development. We then briefly describe the interview process. Chituc (2012) argues that "research on requirements engineering in the context of LTDP is scarce More effort should be allocated to pursue research on requirements engineering targeting information systems ensuring long term preservation of digital data" [1]. From the project's inception DRI emphasised the need to carry out a thorough evaluation of the needs and requirements of its target audience. It sought to underpin the development of this national infrastructure by understanding the activities, task, goals and behaviours of its' users rather than building a solution to an unspecified and unknown problem. To achieve this and to fully understand the problem domain we utilised traditional software engineering techniques and incorporated the use of qualitative interviews in these activities.

Users of the Digital Repository of Ireland can be defined in two ways; firstly users are content holders (cultural institutions, social science archives and libraries) who may either be sharing their digital content directly with DRI or will be sharing their metadata. A second set of users are the researchers and general public who will be making use of the digital content. The boundaries between the two are not clear cut as in some cases, the content holders are also researchers who both archive and use content (for example, the Irish Qualitative Data Archive, the All Ireland Research Observatory, and An Foras Feasa). Additionally the project design includes a number of researcher-led demonstrator projects who are tasked to test the repository and to illustrate the power of the archive. The first round of interviews, which this paper is based on, addressed representatives from the first group of users, the content holders, and the demonstrator projects. Within the interview, the interviewees were asked to describe their needs as both content-holders, and as end-users.

Requirements engineering extracts, derives and specifies system behaviours, operations and functions and ensures the system is built upon and reflects authentic user requirements. DRI considers the problems related to data preservation, manipulation and dissemination associated with the humanities and social science data. This context shapes the DRI solution by ensuring the various user and stakeholder data requirements (e.g. access control) are met. As such one of the most important phases in requirements engineering and development is requirements elicitation, or information gathering. In order to inform the development of the system we need to listen to the target community of users and specify their requirements formally. A user-centric approach, that is listening and learning from the target end-user, is an essential feature of any requirements methodology.

Understanding the problem domain is an essential activity within software engineering and is part of the software life-cycle known as requirements engineering (RE). RE is a subject area in its own right but may also be described as a sub-discipline of software engineering. The RE process informs the development of the system that will be and emphasizes the need for project goals and objectives that are informed by the target audience, or indeed the community of

users. The aim of the RE process is to explicitly state the required features and characteristics of the system from the users' point of view. It is composed of a number of phases, of which elicitation, analysis, specification, verification and evaluation are of significant importance. Our stakeholder interviews were part of the requirements elicitation or information gathering phase. While DRI's mission and vision statement have clear goals and objectives, from which we extracted core business, as well as functional requirements, these interviews highlighted a number of challenges to specifying a clear, generic set of user requirements for DRI.

Policy development is similarly central to the process of becoming a Trustworthy Digital Repository (TDR). The RLG-OCLC Report on 'Trusted digital repositories: Attributes and responsibilities' [2] explicitly identifies policy development as a central function of TDRs. In order to meet these policy obligations, DRI has adopted an eight step policy development cycle;

1. Issue identification
2. Policy analysis
3. Policy instrument development
4. Consultation (which permeates the entire process)
5. Coordination
6. Decision
7. Implementation
8. Evaluation

Part of the policy analysis process required DRI to review national and international practice. Conducting a review, through qualitative interviews with key participants allowed us to not only review policies in existence but to map emerging challenges and areas of concern. This allowed us to develop a richer understanding of policy issues, than if we had limited our review to the collation of published outputs and documentation.

DRI conducted 40 requirement interviews with key stakeholders from December 2011 to August 2012[1]. The representatives were drawn from the following spheres; digital repositories, university libraries, cultural institutions, social researchers, media organisations, public libraries and government content holders. The interviews were semi-structured. Our aim was to establish how users/stakeholders currently support their digital resources/objects and how they develop and maintain their data archives/repositories. The key approach is to use open ended questions (e.g. can you tell me about, can you describe, etc.), following the flow of the interviewee, and only directing, if the issues that need to be discussed do not emerge naturally in course of the conversation.

A topic guide (see appendix) was prepared which addressed the resource/archive in terms of its current data life-cycle.

Pre-ingest Stage: The activities surrounding the data before it is prepared for archiving.

Ingest Stage: Preparation and deposit of data into archive.

Preservation Stage: Fulfilling archive's responsibility to preserve data.

Dissemination Stage: Fulfilling an archive's responsibility to enable reuse of data.

Future development within a federated repository.

Issues addressed included software or computer systems in use, whether it was astatic or living archive, whether there were multilingual data, metadata and database formats, future proofing, data security and user tools. Policy issues relating to ownership, copyright, IP issues, and data sensitivity were also addressed. Where permission was granted, the interviews were recorded and the majority were transcribed. It is intended to archive the interviews so that they will form part of the DRI collection.

3. TRUSTED DIGITAL REPOSITORY - CHALLENGES TO POLICY

Two key discipline-specific policy themes emerged in the course of the interview discussions on facilitating open access. For the humanities and cultural heritage organisations copyright was a key concern, particularly in the face of shifting national and European legal frameworks. Social science organisations required policy frameworks which address data protection needs and the obligation to meet ethical research standards.

Copyright issues were of concern to many. While, there was an eagerness to enable sharing and re-use of digital data, some collections had copyright or ethical restrictions that limited these possibilities. Libraries were affected by the impact of copyright legalisation which placed access restrictions on books, journals and collections they held. Some institutions exercised copyright to generate revenue. Others exercised their copyright in order to limit unwanted re-use of their data. For example, one institution cited the re-use of a photograph in their collection by a commercial entity, in a way that exposed the individuals in the photograph to ridicule. This type of misuse could be prevented by denying the right to re-use. However, this also required that the institution was both aware of the re-use and in a position to defend its copyright – circumstances that would not always be true.

Most social scientific data (and some donations to libraries and archives) had re-use restrictions placed on them which limited who would be able to access the data and required that the anonymity of the original interviewees be maintained. These limitations lessen over time; in 100 years all data can be shared. Our interviewees also expressed concern about long-term preservation of digital content which had time embargoes restricting access, in some cases as long as 30 to 100 years. The time and resources needed to ensure sustainable access to these objects, in order for them to become publicly available in the far future, had not been fully explored by any of our interviewees.

The review found a marked interest in increasing access to digital data, including the use by many institutions of social media to engage with the general public. However there were important tensions. Within the social sciences, where data are collected on the lives of contemporary individuals, a balance needs to be maintained between the rights of the

[1] Ethical approval was granted by the Ethical Committee at National University of Ireland, Maynooth and consent forms (which included consent for future archiving) were used in all interviews.

public to access publicly funded data and the rights of research participants to have their confidentiality protected. Copyright brought with it an additional set of tensions that both restricted the sharing of data and also protected the interests of individuals and institutions. While the copyright concerns attached to digital and physical objects are in many ways similar, digital data carries with it additional opportunities and challenges to make collections and objects widely available by sharing them on the internet, but there was a clear sense that once an object was released, it would then be extremely difficult, if not impossible to police how that object might be used. Given that we are living in an increasingly digital world, there is a need in Ireland for a national digital policy which capitalizes on the possibility attached to digital data and provides guidance on how to facilitate sharing and re-use of digital data. Additionally, internationally a significant trend towards sharing of publicly generated data is evident, and as new copyright and ethical frameworks are developed, barriers against sharing may be reduced. Since the completion of the initial phase of the research, DRI has contributed to the publication of a 'National Principles for Open Access Policy Statement" [3] which in terms of research data states:

Research data should be deposited whenever this is feasible, and linked to associated publications where this is appropriate:

- European and national data protection rules must be taken into account in relation to research data, as well as concerns regarding trade secrets, confidentiality or national security.
- At a minimum, metadata describing research data and its location and access rights should be deposited.

This is an important first step in developing a national infrastructure which facilitates open access and re-use of research data. As such the DRI has adopted an open metadata policy and will make available its metadata under appropriate broad-use licences. It has selected a number of metadata standards which it will recommend for use with textual and visual data and is reviewing metadata standards for other data types. Our decision to support a range of metadata standards requirements is drawn from a recognition, drawn from the interview process, that the various domains served by the DRI have differing experiences in terms of metadata use. Depositors will be advised to use the metadata standard appropriate to their discipline. Our choice of standards reflects common practice in Ireland and internationally[2] Many users are involved with Europeana, therefore an additional policy became evident - the need for interoperability with Europeana. As such EDM will be supported by DRI.

4. BUILDING INFRASTRUCTURE - CHALLENGES TO REQUIREMENTS

A number of issues emerged in the course of the interviews which impacted DRI's requirements specifications; the requirement to retain a local stakeholder identity, clear identification of copyright, variable access controls and the development of user tools, for example time-lines and mapping interfaces. An unexpected outcome was the realisation that as an emerging field many stakeholders did not have a clear understanding of the requirements. The stakeholder consultation was as much a process of discussion as it was of gathering information.

The purpose of the stakeholder interviews was to establish, and learn from, the current activities of the community (relative to the data life-cycle stages mentioned previously) and from this to extract core user, as well as interface, storage and system requirements. Their aim was (and is) to ensure that the system is based on, and supports, *authentic* user requirements. However, through the interview process it became apparent that while we could identify some generic features, there were conflicts and tensions between particular requirements surrounding access, re-use and storage. This is related to the fact that DRI's designated community is quite diverse, both in scope and scale. More worryingly, but perhaps unsurprisingly, many were unclear or unsure of how DRI fitted with their current activities. This created further challenges in extracting requirements (a common problem related to requirements engineering - the customer or user not knowing what they want).

DRI's number one, core, business requirement is that it must be a trusted digital repository (TDR):

> REQ-1 A Trusted Digital Repository.
> The system shall be a trusted digital repository.
> 1.1 It shall supply provide 'reliable, long-term access to managed digital resources to its designated community, now and in the future'. (RLG-OCLC Report). (REQ-34)
> 1.2 It shall conform to the Data Seal of Approval guidelines or equivalent. (Defined by policy).
> 1.3 It shall be an access repository for the humanities and social sciences (HSS).
> 1.4 It shall have disaster recovery process in place. (REQ-57)

This requirement is mandated by the project description and is supported by policy guidelines and decisions. From this high-level, business requirement we specified numerous functional requirements to support the creation of a TDR. These include, but are not limited to, data integrity checks, disaster recovery mechanisms, export functionality and audit trail/reporting. Alongside this it must be an access repository for the humanities and social science data it stores, harvests and aggregates. Our access requirements, that is, how a user or actor can retrieve or view data, state that access to digital objects must be managed through authentication and authorization mechanisms. While we advocate open data and open access it was evident from our interviews that some stakeholders, beyond those with concerns over sensitive data or legislatively imposed embargoes, wanted to maintain some control over data access by particular users. In terms of requirements this solidified the need to implement role based access to content. Conversations about access also raised important questions and concerns over brand identity. Individual institutions expressed anxieties about becoming detached from their own collections within a system such as DRI. This revealed to us an essential user and interface requirement, namely, that of displaying the identity of hosting or contributing institutions or depositors to users

[2] They are Dublin Core, Modified Dublin Core, MARCXML, EAD, MODS and METS

when content was accessed or searched. Alongside this, our stakeholder interviews revealed important issues surrounding data re-use. A key concern was how best to support data aggregation and curation across different collections from different sources without creating copyright and licensing conflicts. Our requirements ensure that copyright statements are displayed to all end users and the systems maps copyright to all digital objects. Access rules foreground all our requirements and specify what a user can and cannot do within the system and in terms of data use and reuse.

5. CONCLUSION

Although the interview process was designed to gather requirements and map and develop policy, it quickly became evident that this was a process of joint discussion between DRI and the stakeholders; interviewees introduced us to the specificity of the issues facing them and we were able to alert interviewees to issues not previously considered. Some features identified are not traditionally seen as part of the remit of a TDR; these tended to be at the level of end-user needs rather than preservation needs (eg. smart phone/tablet use, end-user tools (visualisations, time/maps, user curated collections, crowdsourcing, etc.)). However, the importance of this review process is not necessarily in terms of innovation in terms of data management planning but in creating user-buy in and developing a closer connection between DRI and its user community. In times of decreasing resources and financial pressures (which was a common concern among the community), which creates competition for scarce resources, an approach which develops for community rather than with a community is unlikely to be successful. An "if you built it, they will come" approach is not feasible.

The interviews also highlighted that DRI is unlikely to succeed if seen only as a technical infrastructure. It is a socio-technical system in which the additional roles of training, skill sharing, and national policy development are also central to its mission. Digital archiving was a relatively new field to many; the interviews allowed for mutual learning and fulfilled an unexpected community engagement function. The 'bottom up' approach ensures that DRI will develop in response to stakeholder needs. Policy development continues as an iterative process as both a National Stakeholder Advisory Group and International Stakeholder Advisory Group have been established. Additional stakeholders continue to be interviewed on a rolling basis.

While the interview process has fruitfully contributed to policy development and requirements specification it also alerted us to the necessity for DRI to engage in training and development in order to ensure continued stakeholder engagement with the infrastructure. Building an infrastructure should not be considered a series of linear steps but rather a process of discussion and engagement.

6. REFERENCES

[1] Chituc, C.M. (2012) Requirements Engineering Research and Long Term Digital Preservation Open Research Challenges Workshop, Toronto.

[2] http://www.oclc.org/content/dam/research/activities/trustedrep/repositories.pdf (assessed 24th July 2013).

[3] National Principles for Open Access Policy Statement. http://www.dri.ie/sites/default/files/files/National%20Principles%20on%20Open%20Access%20Policy (assessed 24th July 2013).

7. APPENDIX: TOPIC GUIDE

Stage in Archive Life-cycle	Key Topics	Questions
Pre-ingest	Digital objects/resources? Quantity; data formats (txt, doc), processes of digitisation (crowdsourcing?) Computer or software systems in use. User-interfaces (bespoke, particular product?) Static or living archive? Bi-lingual data?	Can you tell me about your resource/archive/repository? Can you describe your data/content? Is all your data digitised? Can you describe the digitising process? Can you describe the current system you use for your data collection? How do you envisage your resource developing in the future?
	Data Quality Assessment/ Quality Control Process (in terms of data formats and data content)	How do you assess data/content quality?
Ingest	Nature of data (specific concerns, sensitive? rarity, commercial issues). Access issues/policy.	In terms of archiving or storing your data, are there any particular concerns or considerations? How did you address them?
	Ownership/ copyright IP	Who owns the data? Are there copyright issues? Do you have licensing agreements? Are there any IP issues?
	Collection priorities.	How do you source the data? Do you have specific priorities?
	Catalogue Ontology/ Thesaurus	Have you developed a catalogue? If so, can you describe it?
	Metadata formats? Database formats? Linked Data? Open Data/	What metadata standards do you use? Would you know what the database system you are using is? (MySQL, Excel, XML etc.)?
Preservation	Future-proofing - data formats/longevity of data.	Can you describe your preservation process, if any?
	Data security (physical threats, virtual threats) /Redundancy	Where is the data physically stored? What security systems do you have in place if any?
Dissemination/Data Re-use	User Experience /expectations (Actors e.g. students, researchers etc.).	Can you describe who uses your data? How do you see users in the future?
	What tools etc. do users currently use? (bespoke or not)	Do you provide any tools to enable the user to interact with the data?

Archives New Zealand Migration from Fedora Commons to the Rosetta Digital Preservation System

Jan Hutař
Digital Continuity
Government Digital Archive
Programme
Archives New Zealand
Te Tari Taiwhenua
Wellington, New Zealand
jan.hutar@dia.govt.nz

ABSTRACT
This paper discusses the New Zealand Government Digital Archive Programme (GDAP) and its requirement for Archives New Zealand to move to a fully functional digital preservation system. It looks at the migration of digital content from Fedora Commons to Ex Libris' Rosetta Digital Preservation System focusing on what needed to be migrated, preparation of the migration, how it was performed and what tools were needed to support the work. We look at the verification of this process and conclude with an audit of the results and a description of the lessons learned during this process.

Categories and Subject Descriptors
H.3.7 [**Digital Libraries**]: Systems Issues; J.1 [**Administrative Data Processing**]: Government

General Terms
Management, Measurement, Verification.

Keywords
Digital Archive, Migration, File Formats.

1. INTRODUCTION
Archives New Zealand, *Te Rua Mahara o te Kāwanatanga*, bound by the public records act 2005 **[1]** is the sole keeper of the memory of government in New Zealand. In 2010 it was announced by government that $12.6 million (NZD) would be made available to fund a Government Digital Archive (GDA), in order to improve management of the increasing number of digital records created by public sector agencies.

In support of this initiative, $9.7 million (NZD) was allocated to Archives New Zealand for the GDA across four years. $2.9 million (NZD) was allocated to the National Library of New Zealand for its equivalent programme, the National Digital Heritage Archive (NDHA). The goal was to work on the GDA project in co-operation with the library **[2]** by utilizing its existing systems. The NDHA programme started in 2005 and went live with its first digital preservation system in 2008.

Before GDAP, Archives New Zealand had the Digital Continuity Action Plan - endorsed by Cabinet in 2009. Building a robust digital archive system and processes is crucial to fulfilling statutory responsibilities for the long term preservation and accessibility of digital data from agencies. The GDA programme is one of the main outcomes of the Digital Continuity Action Plan.

As a part of the streamlining of government agency structures, The National Library of New Zealand and Archives New Zealand were formally incorporated into the Department of Internal Affairs (DIA) on February 2011. One of the consequences of this change was the decision that the GDA would leverage from previous government investment and research by sharing the existing digital preservation system of the NDHA – Rosetta **[3]**. Using one system required extending existing infrastructure, including hardware, architecture and the long-term preservation system settings, as well as the development of additional software capability. Another consideration is the development and understanding of organizational responsibilities and processes, as well as the creation of shared policies across both institutions.

The move to utilize the same systems developed by the NDHA required the migration of the content from Archives New Zealand's Interim Digital Archive (IDA), built on top of Fedora Commons, to Ex Libris' Rosetta digital preservation technology; this also required the integration of Archive New Zealand's "catalogue, collections management and public search" system - Archway. Migration of the IDA content is the first of the three releases planned as part of GDAP.

It is hoped that the new infrastructure will help Archives New Zealand to achieve five principle objectives **[4]**:

1. Protect important public sector digital information through change
2. Empower government, businesses, and communities to discover, access, understand, and reuse important public sector digital information
3. Foster digital continuity understanding with stakeholders
4. Streamline the transfer of information from public sector agencies to Archives New Zealand
5. Support the public sector to achieve the purposes of the Public Records Act 2005

2. REPOSITORY MIGRATION
2.1 Interim Digital Archive - Fedora Commons
Fedora Commons was selected and implemented in 2008 to provide Archives New Zealand with an Interim Digital Archive (IDA) after the establishment of the Digital Sustainability Programme. With longer term planning already underway for a programme to implement a complete digital preservation system,

IDA provided the organization with digital repository functionality that could potentially be replaced within 2-3 years. Active preservation of existing, archived digital materials was considered low priority for the IDA. Archives New Zealand identified several benefits of Fedora as a short-term solution for a digital repository including:

- Zero proprietary product costs and constraints
- Customizations being easier to make due to open source code base
- Fine grained security; support for up to one million digital objects
- The advantage that one other New Zealand government agency was using it - the State Services Commission, *Te Komihana O Ngā Tari Kāwanatanga* (SSC)

It was clear from the beginning that the Fedora based IDA would provide just the minimum functionality to support the business processes involved in accepting and managing a digital archive, that is, the ability to ingest data, manage archival objects and provide access to them via Archway. We knew it had limited functionality to support complex digital preservation. It was also necessary to build the IDA for the increasing number of materials being digitized for access. It was never used for storing data from physical carriers like floppy discs, CD/DVDs etc. which Archives New Zealand received from a handful of agencies. No digital transfer has ever been ingested into Fedora though it was one of the reasons for establishing it. The ability to accept digital transfers is one of the main deliverables of GDAP.

Fedora provided an adequate solution for an interim digital repository, but the technological infrastructure it was established on was limited. It was not a system that could be scaled to provide a 'whole-of-government' solution. The requirements of GDAP demanded more robust hardware and a logical digital preservation solution. So the decision was taken to align Archives New Zealand's technical approach for a digital repository with the well-established repository maintained by the National Library of New Zealand.

2.2 Rosetta

Rosetta is a long-term preservation system developed by Ex Libris. It may be considered an outcome of the NDHA programme, where initial requirements for such system began taking shape in 2005. This was originally in partnership with Endeavor Information Systems (Elsevier), later taken over by Ex Libris, who became the primary partner for developing the software package (after acquiring Endeavor in 2006 **[5]**). Rosetta has been used as a digital preservation system at the National Library of New Zealand since 2008, when its first version went live with the launch of the NDHA digital archive.

Presently, both institutions are using a shared implementation of Rosetta 3.1. There are currently 17 customers of this system around the world [Email communication with Nir Sherwinter (Ex Libris) on 4 April 2013].

2.3 Process

In preparation for the migration we needed to get details of the IDA content. The repository contained about 40TB of data at the initial stage of migration planning in late 2011. This became 48TB as data was ingested into the IDA during 2012, until the new digital repository was switched on in December 2012. The IDA mainly stored digitized documents from collections like the personnel records of First World War soldiers from the New Zealand Defence Force (NZDF); Westland maps; Land Information New Zealand (LINZ); the Treaty of Waitangi and other collections. Each collection had been appraised regarding the importance of the documents; the necessity of migration, that is, if they were already linked with Archway; and the difficulties expected in a migration. There was an Excel spreadsheet for each collection listing items' ID, title, description, collection ID, and the reason for migration or for leaving it out of the process.

Rosetta can ingest data in a certain shape and structure and with a certain metadata format. The Rosetta data model is based on the METS and PREMIS standards. Every Submission Information Package (SIP) ingested into the system has to be wrapped in METS with DNX metadata. DNX is Rosetta's proprietary metadata standard which can contain PREMIS-like metadata among technical metadata standards such as MIX.

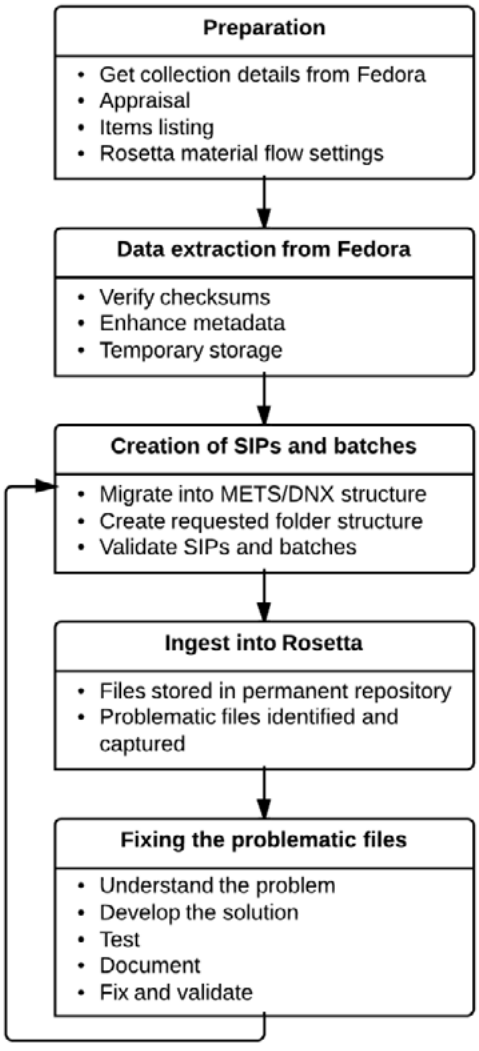
The First step of the data migration was extracting the digital objects and metadata from the Fedora repository. Objects were stored in Fedora with minimal metadata, because the descriptive metadata is stored in the Archway database, which is the archival description management system developed and used by Archives New Zealand. Key metadata for the migration was the Archway ID and checksum values. The Archway ID is used for linking between data in Rosetta and their description in Archway. If an MD5 or SHA-1 checksum was present for a file in Fedora, it was re-calculated after the file was extracted and compared against the stored value. Warnings were produced for:

- missing checksums,
- unsupported checksum types,
- failed checksum checks resulting in a failure for the item.

The next step was to migrate extracted objects and metadata into a new structure to comply with the Rosetta internal metadata standard and put them in the temporary storage location. SIPs can contain one or many items, but after testing it was decided to maintain a 1:1 ratio, that is, **one SIP for one item or record**. Each SIP has a METS structure and basic descriptive Dublin Core metadata (Title, Provenance, Series Number and Archway ID).

Rosetta will accept an item for deposit when packaged as part of a SIP. Generated SIPs were sorted by collection and arranged into suitably sized 'ingest batches' for overnight processing. Each batch of SIPs was prepared for pre-ingest, to check for zero-byte files, non-existent METS, and exceptionally long or improperly formatted filenames. We also checked for duplicate items, both within the same batch and against previously ingested ones. Finally, we triggered the ingest process in the Rosetta deposit module – see in the Figure 1 below.

Figure 1: General workflow of the migration process



2.4 Migration tools and Rosetta settings

For the data extraction from the IDA, its preparation and migration into the new SIP structure, a java based "Migration Tool" was developed. The tool recurses the IDA repository looking for all items with a published Archway ID and then takes the data from Fedora repository and puts it into a temporary location. The tool also invokes the process of creating the SIP in the structure Rosetta expects to receive, that is, with METS structure and additional administrative metadata (date of ingest, agent etc.).

Scripts were created to generate the batch for ingest, validate the batch and trigger the ingest itself via the Rosetta deposit API.

At the moment, we have a Producer entity (see below) in Rosetta based on the method of ingest. This means that for the automatic ingest of migrated data batches we created a new Producer called Archives New Zealand Digital Migration (ANZDM), in addition to the existing Archives New Zealand Internal Digitisation Programme (ANZIDP), which is used for all data coming from the Ingestor User Interface (UI) application. The Ingestor application is used by our archivists for ingesting digitized data on a daily basis. Each Producer in Rosetta has one or more agents that represent real actors from across Archives New Zealand, who are allowed to ingest if they have proper Rosetta and Ingestor roles and access rights. A Producer can be connected to one or more ingest material flows and could have different requirements for the metadata provided and file formats permitted for ingest etc.

Settings for the Fedora migration ingest material flow in Rosetta were the same as that for the normal ingest of digitized data via the Ingestor UI application. The steps of this ingest material flow include file format validation, virus check, risk assessment, structure validation, metadata extraction and access copy generation (for example TIFF to JPG). There were different material flows for different file collections; the only difference being the specifications used for access copies. For example, JPEG at 900x900 px resolution for NZDF A4 format documents compared to 3000x3000 px for maps. The goal was to avoid any unnecessary manual intervention. The only point where manual work is required is where files fail technical assessment, for example via DROID or JHOVE characterization, and end up in the Rosetta Technical Analyst Workbench for further investigation by a Digital Preservation Analyst.

The results, such as the number of ingested items and files, of each batch ingest could be confirmed by querying the Oracle database of Rosetta.

2.5 Result

There has been a total of 70 bulk ingests run over a period of 12 months - batch 001 on 7 February 2012 and batch 070 completed on 20 January 2013. The estimated total amount of data in the IDA was calculated at 48TB. On completion approximately 46TB (45,9TB) of data had been migrated, which represents 63,460 archival items (not files, item has 30 files on average). All of the items have been extracted, ingested into Rosetta and synchronized with the access portal Archway. The rest of the original 48TB were not migrated as they were not associated with any existing Archway item or temporary usage.

Items that were not successfully migrated into Rosetta, usually due to some technical problem associated with one or more file streams, were moved to a "quarantine" directory, which was kept on a NAS storage device. In total 453GB of data (0,1%), which equates to 468 items / SIPs have been quarantined.

Migration was scheduled across one year. On average, for each migration we ingested batches of 723GB in size. The biggest ingest was 1,2TB. Each batch consisted of anything between 10 to 5,000 items depending on the type of material being ingested. The limitation on size came from the expected ingest time required to process each batch. It was necessary for it to complete before 8am each day because we did not want this process running during normal working hours when it may impact on other processes. The average time taken for an ingest was seven hours. Performance was dependant on hardware configuration, volume of files and their file size. We were able to use our current hardware configuration with this number of batches and files; there was no requirement to complete ingests faster and therefore no need to upgrade the hardware for the Rosetta deposit module.

Another reason for restricting batch sizes was because of the time involved in auditing and reporting the results each day following the process. Problems that had occurred during a bulk ingest had to be resolved prior to preparation of the next ingest. It was only

possible to prepare the next batch on completion of the last, due to the risk of including undetected duplicate objects.

Our final audit was done via the Archway database, where all items are stored. It was compared with the original list of item IDs stored in IDA repository and then with the current Rosetta Oracle database of item IDs in our production environment. Put simply, if an IDA item has an associated Rosetta item ID, we can say that it has been synchronised with Archway via the Rosetta publishing process and therefore successfully migrated. We have identified only two duplicate items ingested during the entire operation.

3. FILE FORMATS

As mentioned, almost 0,5TB of data was identified as problematic and moved to quarantine outside of the Rosetta system. All of the issues related to problems with the bit-streams of files themselves, format identification, validation, or subsequent metadata extraction from these files. Tools like JHOVE generally will not validate files which do not conform to the specification of the format. Therefore in some cases no technical metadata is created, which is a major problem for us as we aim to create and keep as much technical and administrative metadata as possible. Also, we aim to have consistent metadata for similar file types. The IDA did not provide file format validation, identification and metadata extraction on ingest. No quality assurance on file formatting, validity, or structure was done, either for internal digitization outputs, or for digitized data received from external digitization companies. We have migrated only digitized documents and for that reason the file formats were limited only to TIFF files and PDF files.

3.1 General overview

The table below shows the list of issues we encountered ingesting digital objects from the IDA into Rosetta. All files were caught in the Rosetta Technical Analyst Workbench. In this environment the Technical Analyst can perform a technical assessment of each file and understand what is causing the issues, for example by looking at the JHOVE validation output, or messages from other tools. It is also possible to solve the issue; for example, in the case of multiple file format identification in DROID, choose the right identification; or in other cases download the file, investigate and fix the problems and upload the fixed file back into the workbench to be sent to the permanent archive after it is re-validated and relevant metadata extracted.

Table 1: List of issues encountered during ingest into Rosetta

	Error Message	File Format	# of SIPs
1	Tag 305 out of sequence	TIFF	197
2	Tag 270 out of sequence, Tag 269 out of sequence	TIFF	112
3	Invalid ID in Trailer	PDF	94
4	Exception occurred during metadata extraction	PDF	41
5	Unknown field with tag 347 (0x15b) encountered. Invalid YCbCr subsampling. Cannot handle zero strip size missing an image filename	TIFF	30
6	Invalid DateTime separator	TIFF	23
7	Multiple formats found for file	TIFF	4
8	Checksum Error, Premature EOF	TIFF	2
9	Malformed dictionary: Vector must contain an even number of objects, but has 3	PDF	2
10	Count mismatch for tag 36864; expecting 4, saw 0	TIFF	1
11	Improperly nested array delimiters	PDF	1
12	Invalid character in hex string	PDF	1
13	Invalid page tree node	PDF	1
14	Invalid strip offset, JHOVE message: Invalid strip offset, Invalid DateTime separator: 2010/09/28 02:39:27	TIFF	1

In the process of migrating data from the IDA into Rosetta, we chose an approach more suitable for large amounts of data; we did not try to solve all issues in the Rosetta Technical Analyst Workbench, rather we moved all SIPs caught in the Technical Analyst Workbench to our own quarantine location. There, the digital objects were analysed, fixed in bulk with an agreed solution, and the whole SIP re-submitted into Rosetta. This allows Archives New Zealand to avoid too many individual issues sitting in the Technical Workbench to be resolved and to allow us to continue with the remainder of the migration process.

3.2 Issues

Error messages in Table 1 are mainly output by JHOVE. Issue 5 is from ImageMagick, which is used in Rosetta for creating JPG access copies from TIFF masters. If the creation of access copies is not done for a file, the whole SIP is routed to the Technical Analyst Workbench again. Below is short description of some of the issues we encountered.

The most frequent issue was related to the bad formatting of files, in particular TIFF files with their tags out of sequence. The error output *"Tag 305 Out of Sequence"* from JHOVE is a little misleading, in that the tag is not strictly out of sequence. The problem is that there are two TIFF 305 'Software' tags in the metadata, each containing a unique string value. Only one Software tag is permitted in the TIFF specification. This was a problem generated by one of scanners used by the digitization company that created these files.

A similar problem with "Tag 270 out of sequence, Tag 269 out of sequence" appeared in 112 SIP packages. This related to TIFF metadata, tag 270 ImageDescription and tag 269 DocumentName, which were populated accidentally by the digitization company.

The error Invalid ID trailer[1] in our PDF files was created because we merged two PDF files into one in our workflow for creating multipage PDF access copies. That is, PDF file containing all the scanned pages of a certain file was combined with a PDF cover

[1] ID entry is an array of two byte-strings constituting a file identifier for the file. File identifiers are defined by the optional ID entry in a PDF file's trailer dictionary **[6]**.

page with relevant information about the original file (Archway ID, Title etc.). The issue with the PDF trailer was caused by the PDF creation engine Multivalent used in our environment.

The ImageMagick error showing: "Unknown field with tag 347 (0x15b) encountered. Invalid YCbCr subsampling. Cannot handle zero strip size missing an image filename" was caused by the appearance of non-standard features of some TIFF files (JPEG compression in TIFF, PhotometricInterpretation TIFF baseline tag with YCbCr value etc). Again, this was different to other TIFF files from the same collection and was a processing error during the digitization and post processing. We also discovered that only libtiff v3.9.4 of ImageMagick had problems handling those TIFF files, previous and later versions of libtiff worked fine.

The final issue we should highlight was that of poorly formatted and thus invalid date time separators in TIFF baseline metadata tag 306 DateTime. The TIFF format standard **[7]** specifies that this value should be formatted as [YYYY:MM:DD HH:MM:SS], whereas all the ingested LINZ images had a "/" (forward slash) instead of a ":" colon, that is: [YYYY/MM/DD HH:MM:SS].

3.3 Common solutions

In order to complete the migration, we had to solve all of the issues and re-ingest the data into Rosetta. While it is possible to ignore errors, Archives New Zealand's policy is to deal with problems when they appear. Ignoring the problem would very likely cause other problems in the future, for example while trying to complete a preservation migration of poorly-formed file types, such as our TIFF examples, into a new preferred file format. We would not consider the list of issues serious and they are unlikely to cause problems rendering the file. Each file in the above examples could be rendered, but not always technical metadata was created by metadata extractors because the files were incorrectly formatted. If we were to ignore this it would mean that we have in our permanent archive some preservation master TIFF files with technical metadata and some without. This inconsistency could limit our ability to access and work with these files in future; for example, the ability to search based on metadata fields and then create sets of documents with certain features, or more importantly, to assess the risk linked to certain files and formats.

The majority of the issues were fixed with scripts developed in-house. These were sometimes very basic and might simply call relevant tools like ExifTool for changing the metadata. Each problem and its solution were thoroughly analyzed, tested and documented. The aim was to introduce as minimal a change as possible into the bit-stream of each of the relevant files. For analysis of erroneous files we used community standard tools such as JHOVE, DROID, NLNZ Metadata Extractor, FITS and basic hex editors.

Each issue has been thoroughly documented and that documentation has been saved in the Archives New Zealand EDRMS. EDRMS IDs of the documentation files were then added into the metadata of the corrected digital objects. The documentation consists of the problem description with links to the relevant file format documentation. There is a list of options for dealing with the problem and finally the decision about the preferred solution. Another part of documentation is about how the solution was tested. Custom scripts are also stored in the EDRMS. The idea behind this is that all changes to files have to be documented and referenced from the item metadata, so that future users can understand what was done and why.

All this is considered to be pre-conditioning and follows Archives New Zealand's Pre-conditioning Policy which was developed alongside the National Library of New Zealand. The Pre-conditioning Policy deals specifically with changes to digital content that has come within the control of the Archives or Library, but has not yet been ingested into the preservation system. It focuses on objects where there is a need to solve technical issues. The policy covers changes to content that do not result in both the original file and a copy being ingested. Pre-conditioning changes are made entirely on the original - they do not generate a new copy[2].

Conditions for pre-conditioning in the policy are that the nature of the change is completely reversible and not extensive; it cannot change the intellectual content and it must be documented. The preservation system must also store a provenance note. This note should remain as part of the file's preservation metadata throughout its existence. This is true for all changes of the digital objects and resulting metadata mentioned above.

The provenance note is automatically added into the metadata as an event of the preconditioning. It consists of a short description, the outcome and a reference to the documentation in the EDRMS. This process is part of each script used for solving the aforementioned issues.

4. CONCLUSION

Migrating 46TB of data is a big task. One would hope a minimal set of issues are likely to arise. To ensure a smooth migration, there are a couple of steps that need to be completed before the process begins. There has to be a plan, an analysis of current content, an ability to deal with issues and a mechanism for audit at the end. Handling issues as they occur and before ingest might prove to be the most time consuming part of the whole migration but ultimately makes the files more predictable to handle the next time around. In our case we had policies in place that helped speed up the decision making about what to do about different issues and these will continue to assist us in the future.

We have learned a lot from this migration. First of all, very few issues came from the migration itself. There was no lost or corrupted data. The main issue was the quality of the data, which had not been checked before that point. Data and file formats were not validated in the IDA solution. A key learning is that we now plan to do basic validation, with tools like JHOVE and DROID, as part of Archives New Zealand's internal process before accepting digitized data from external vendors.

Migration also helped us to understand the nature of the problems we will have to face once we start transfers of born-digital content from government agencies. Our approach to fix all the issues and keep as consistent an archive as possible might prove to be unrealistic while trying to cope with the flood of different types of digital objects from transfers.

We were also pleased to see that the Rosetta digital preservation system could easily cope with 1TB of data ingest in 6-8 hours

[2] If this is the case, its covered by the Preservation Action Policy and such a change must happen in the controlled environment of the preservation system.

within our current infrastructure. It was confirmation of our early expectations.

The last step of the migration was a final audit of the data in Rosetta and deletion of the Interim Digital Archive content - this was completed in July 2013.

5. ACKNOWLEDGMENTS

My thanks to colleagues Mike Ames, Ross Spencer, Tracie Almond, Matt Painter and GDAP manager Alison Fleming for being able to use internal documents which they have created.

6. REFERENCES

[1] Department of Internal Affairs of New Zealand. 2005. *Public Records Act*. Public Act 2005 No 40. http://www.legislation.govt.nz/act/public/2005/0040/latest/DLM345529.html

[2] New Zealand Government. 2010. *Announcement of Government Digital Archive* [online]. [cit. 20-04-2013]. http://www.beehive.govt.nz/speech/announcement-government-digital-archive

[3] Hutař, J. 2012. Assessing Digital Preservation Strategies. In *International Council on Archives Congress* (Brisbane, August 20 -24, 2012). http://www.ica2012.com/files/pdf/Full%20papers%20upload/ica12Final00155.pdf

[4] Archives New Zealand. 2010. *Government Digital Archive Programme* [online]. [cit. 20-04-2013]. http://archives.govt.nz/advice/government-digital-archive-programme

[5] Elsevier. 2006. *Francisco Partners to Acquire Endeavor Information Systems from Elsevier* [online]. [cit. 20-04-2013]. http://www.elsevier.com/about/press-releases/science-and-technology/francisco-partners-to-acquire-endeavor-information-systems-from-elsevier

[6] Adobe Systems Incorporated. 2000. *Adobe portable document format*. v 1.3. 2nd ed. p. 477. Addison-Wesley: Boston. ISBN 0-201-61588-6. http://wwwimages.adobe.com/www.adobe.com/content/dam/Adobe/en/devnet/pdf/pdfs/pdf_reference_archives/PDFReference13.pdf

[7] Aware systems. 2008. *TIFF Tag DateTime* [online]. [cit. 20-04-2013]. http://www.awaresystems.be/imaging/tiff/tifftags/datetime.html

Destination: Shared Repository
The National Library of France's Journey to Third-Party Archiving

Louise Fauduet
Department of Preservation and Conservation
Bibliothèque nationale de France
Quai François Mauriac
75706 Paris Cedex 13
louise.fauduet@bnf.fr

Sébastien Peyrard
Department of Bibliographic and Digital Information
Bibliothèque nationale de France
Quai François Mauriac
75706 Paris Cedex 13
sebastien.peyrard@bnf.fr

ABSTRACT
The SPAR repository project started as a way to make preservation easier, cheaper, and more effective for the National Library of France (Bibliothèque nationale de France, BnF). At the time (early 2000s), the BnF used different storage media and technologies across the library, and had no unified responsibilities and processes. When the decision was made to overhaul the infrastructure and the software to have one repository to replace them all, the project designers reflected quite naturally that such an effort could and should benefit other libraries that had similar scalability issues.

There were many obstacles to overcome after that generous impulse. First, finding the BnF's niche in the digital preservation landscape, at a time when it was growing and evolving fast — as it still is. Even when focusing on the heritage sector, there are several other repositories or planned systems that have a national vocation, within the archives or higher education communities for instance. Then came the matter of combining the needs of the library itself and the design necessities of third-party archiving to create a repository that would be scalable, trustworthy and open. Last but not least, the BnF is continuously refining the way the repository can best serve its clients. The most accessible function for partners is bit-level preservation, and any extra step toward comprehensive preservation has to be balanced with available resources and tiered prices. As the first clients come in, prices and processes are still in flux, and vulnerable to policy changes.

Keywords
Digital repository. Third-party archiving. Cost models.

1. INTRODUCTION: ARCHIVING AS A LIBRARY?!

1.1 What we talk about when we talk about archiving

The French language is fraught with nuances that do not quite translate into English. One such problematic word is "archive": amongst cultural heritage professional circles, archives are first and foremost the place where records go when they grow up and become "permanent", while the French word "*archive*" is applicable to any stage of the archiving process. And archives are not the libraries' domain, usually, except when authors' papers lose their way and enter library collections. What's more, in France, the term "*archivage*" can be used for records management, archiving and preservation, or third-party archiving where it is called "*tiers archivage*". In other terms, the same "archive" term can be be used for curation or for preservation, depending on the context. The same problem occurs with "tiers archivage", which means third-party archiving but can be applied for preservation as well as archiving services. Last but not least, a term like "Web archiving" is used to define something which, in France, comes under Legal Deposit law.

In other terms, "*archivage*" can be applied to different professions, skills and activities (archives, libraries), and to different legal statuses (administrative production or publication). Moreover, when it comes to third-party preservation of archive records, any institution entitled by the central services for French archives to perform such activities, can paradoxically do archive preservation without being an archive center in the first place[1].

France's deposit and archiving laws may be puzzling as well. Certain types of materials have clear destinations in paper and digital forms, others are not constrained or are unclear. For instance, the publication of a research center in a University can be considered a publication and thus subject to legal deposit, but can also be considered a public archive record as a product of an agent of the state. Which legal system applies to it depends very much on the heritage institution that will take on the responsibility of preservation

1.2 France's digital preservation landscape

If we focus on the cultural heritage sector, France's digital preservation landscape is shared between a few large institutions, especially in the Ministry of Culture (public libraries and archives) and the ministry of Higher Education and Research (universities, research centers and datacenters). The administrative distinction between those two ministries is very relevant in understanding why the three main systems in France involved in heritage digital preservation are the BnF's SPAR system, CINES' PAC platform and IN2P3's datacenter for scientific experimental data. This rather centralized landscape fits the recommendations of the report called "Strategic orientations for digital libraries"[2], produced by the president of the BnF under the auspices of the Ministry of Culture in early 2010, which insisted upon lowering public costs in order to produce economy of scale. On the archives side, the landscape is more parceled out: the French Archives ministerial services have a role of technical recommendation, control and advice, but some local archive centers have developed

[1] This is the case of the CINES and BnF who, among other organizations, received the grant to ensure archive preservation.

[2] Called in French "Schéma Numérique des bibliothèques": http://www.enssib.fr/bibliotheque-numerique/document-48219.

their own solutions. The VITAM[3] project intends to provide a large scale solution for the National Archives and the Archives of the Ministry of Foreign Affairs in a three-year time frame. Another solution called CDC Arkhinéo[4], targeted at third-party legal archiving, has been developed by the French public institution called Deposits and Consignments Fund ("Caisse des dépôts et consignations") with a strong focus on security and legal evidence. The National Center for Scientific Research has a solution for digital humanities, Huma-Num[5]. At local scale, some repository solutions are being developed, e.g. M@rine developed by two department archive centers[6] And several private firms, some with experience in records management and archiving for banks, for instance, are offering their products to public archives.

In compliance with the strategic orientations mentioned above, the DISIC (Interministerial direction for IT Systems) created in 2011, has a similar mandate of rationalizing the public expenses on IT infrastructures. Its focus on digital preservation, however, will only focus on technical aspects, leaving the key organizational challenges outside its perimeter.

Given the history and context of the SPAR project, the BnF services are somewhat different from the most common shared repository models:

1. Projects that started with a national or local mandate and *ad hoc* governance structure, developing and sustaining the repository for its members (National Digital Library of Finland, HathiTrust), with a partnership model;
2. Projects with a national or local mandate to provide a service to a community, where the service provider has no collections of its own in the repository, and the customers are not part of the board (CINES, California Digital Library's Merritt);
3. Software solutions where the vendor fosters a community of users, either as a downloadable software (e.g. SDB, Archivematica…) or as an online facility (Duracloud, Preservica…);
4. Networks of repositories exchanging copies of their information packages (e.g. LOCKSS networks, Chronopolis, TIPR…)[7].

The BnF sells storage and services, but mostly maintains control over the technical roadmap and the repository governance, as SPAR was first developed for its own preservation needs. So far it is closer to an institutional repository model.

[3] http://www.archivesnationales.culture.gouv.fr/chan/chan/english-version-colloque-archiving-2013.html.

[4] http://www.cdcarkhineo.com.

[5] http://www.tge-adonis.fr.

[6] This solution ensures preservation as well as archive-specific curation functions. Cf. http://www.sicem.fr/index.php?option=com_content&view=article&id=167&Itemid=41.

[7] A recent census of existing preservation repository initiatives can be found in Aligning National Approaches to Digital Preservation Conference Edited Volume. http://educopia.org/sites/educopia.org/files/ANADP_Educopia_2012.pdf.

2. MAKING A SHAREABLE REPOSITORY

2.1 Looking for scale

The BnF's main strength compared to other heritage institutions is the size of its own collections. SPAR became operational in May, 2010. As of June 2013, the repository hosts around 1 million information packages, representing over 800 Tb, essentially from the library's digitized collections. Many more hundreds of terabytes from the backlog of digitized collections and from Web archives collections are being ingested. The current storage capacity of the system is 1 Pb, with about 16 times more in terms of slots available in the tape library. This may not be very sizable on the international scale, but it is for example much more than the CINES has budgeted so far for the collections its repository stores, at 40 Tb. There is no doubt that the other repositories will grow, but the BnF's SPAR has a head start given the library's own needs, and thus has already achieved a certain economy of scale regarding storage.

What's more, the software itself has been designed to scale up. By making it as modular as possible, the development team hopes to be able to change any given module (ingest, storage, data management, etc.) according to new progress in technology or new requirements in scale. Another strategy has been to add multiple instances of the most used modules, which deal with SIP preparation and ingest.

Above all, the design for SPAR has been based on the concept of "tracks" and "channels", to organize content and make managing heterogeneous collections easier. Tracks are created according to the legal status and entry mode of the collections they enclose: digitized materials, legal deposit, gifts and acquisitions, etc. A track for third-party archiving has been envisioned from the beginning. Channels are sub-divisions of tracks according to technical challenges and refinements in preservation requirements. With each channel, a new set of service level agreements are negotiated, defining the conditions for ingest, preservation and access.

Thus the logic of the system is to have at least one new channel created for each third party submitting assets to the SPAR repository, with its own set of negotiated parameters. Should the nature of the collections entrusted to the BnF by a third party be varied in its technical composition or in the level of care it requires, then more channels should be added. The upside to this is a high adherence to the needs of the partners; the downside is the extra burden on the BnF's staff and resources each time a new channel must be set up and maintained.

2.2 Looking for standards

Making the philosophy and design of a repository compliant with standards is a key condition to its being shareable. Hence the use of the OAIS standard to design the system and the use of the METS and PREMIS standards for its data model and the preservation metadata of each document. These proved to be invaluable since initiatives can be initiated on the international scale that benefit back to the repository. For instance, the BnF will take part in the Preservation Health-Check Pilot[8] in the course of 2013, whose purpose is to give a risk driven evaluation of the METS/PREMIS metadata stored in the SPAR repository. The BnF could not have been part of such an international R&D projet without standard metadata formats. Another great added value those standards provided was genericity, whatever the kind of

[8] http://www.oclc.org/research/activities/phc.html.

content was; and, in a longer-term perspective, lower the barrier to making the other systems and initiatives mentioned in 1.2 interoperable with the BnF repository solution to allow distributed preservation over the country.

In addition, efforts were made to use open source software whenever possible in SPAR's own code. The principles are the same as with the use of standards: benefiting from community-approved tools, and adding to them whenever possible (the BnF has commissioned an ARC and a GZIP module for JHOVE 2, for instance), while fostering interoperability.

2.3 Looking for certification

Once the BnF decided to open its services to third parties, it was important to prove its trustworthiness to them. To this end, the BnF has been monitoring the certification initiatives that have started ever since the OAIS was first published, including the TRAC and DRAMBORA check-lists. With the birth of the European Framework for Audit and Certification of Digital Repositories, in 2010, the path to certification is now clearer. However, on top of international certification, the BnF is also concerned with French standards and certifications.

It is currently interested in three 3 parallel certification initiatives:

1. The authorization to preserve third-party archive records, required by French Law for an institution or a firm to be entitled to store and preserve public administrative documents;
2. The French AFNOR[9] Z42-013 standard, which evaluates the technical trustworthiness, security and traceability of the preservation system. It has been transformed into ISO-14641-1 at the international level. (ISO 14641-1:2012, Electronic archiving -- Part 1: Specifications concerning the design and the operation of an information system for electronic information preservation[10]). The corresponding French certification was created as NF-461 in early 2013;
3. The Data Seal of Approval[11], which evaluates the OAIS compliance of the repository

The BnF received the first of these in Spring 2013, after a year of discussions with the central services for the French Archives at the Ministry of Culture, who deliver the authorization.

These efforts have revealed two main issues with the BnF's certification efforts, that will be addressed in the coming months with the help of the person in charge of disaster and risk management at the library, and with new software developments:

- the lack of policy statements at the library level – the preservation policies have so far been discussed and implemented with collection managers directly – and of technical documentation in English. Those documents would be essential to getting a Data Seal of Approval, for instance;
- the low level of security required for the preservation of the library's own collections, compared with the authenticity standards expected in dealing with public or private archives, for instance, due to their potential roles in judicial processes.

Thus, while working on its digitization or Web archives collections, or even on the initial phases of its third-party archiving services, the BnF hasn't invested in time-stamps, tamper-proof hashes, and certified signatures. It is working on these aspects now that the third-party services are attracting more interest.

3. GETTING CLIENTS

3.1 Defining services

The SPAR system has been developed in an iterative way: after the core functions were created in 2008-2009, and after a first track was set up for digitized collections, seen as the most urgent preservation need at the time, the team went immediately on to work on the third-party archiving track.

There were managerial motivations to this decision, as the push for unified repositories at the State level was already being felt. There had also been a trend within the library to generate income from its own services.

For the system designers, this represented an opportunity to make it more generic and customizable. The core of the repository is intended to be as generic as possible: all functions dealing with SIPs, AIPs and DIPs must be available to all tracks and channels, whether they use them or not, so that the ingest, data management, storage and access modules are standardized and easier to maintain. However, to deal with the wide variety of objects to be preserved at the library, specialized pre-ingest modules are created for each channel, in order to turn specific information submitted by the producers into normalized SIPs.

The first pre-ingest module created for SPAR dealt with the highly specific requirements of the BnF's history of digitizing its own collections according to the strict rules dictated by preservation needs, but also by the constraints of the BnF's digital library, Gallica. Working on third-party archiving meant building bricks for pre-ingest that were as simple and as universal as possible in order to make a nonetheless acceptable SIP. The focus was thus on bit-level preservation, with an added service of metadata processing. The client can submit metadata files with its information packages, and the metadata will be mapped through an XSLT file to the descriptive metadata section of the SIP's METS file. The rules for detecting these metadata files and the content of the XSLT are entered into the service level agreement.

The idea was that after a first phase of experimenting with bit-level preservation, upgrades to the third-party archiving track would be designed in accordance with actual clients' needs and requirements.

The internal benefits of developing software for non-BnF communities were not negligible: the repository now had a redesigned pre-ingest architecture, that relied on common functions, and a model for simple, versatile pre-ingest modules which could be re-used for preserving library collections when time, resources or maturity meant that advanced pre-ingest functions could not or should not be developed. It has been used to deal with the digital versions of advertising posters, for instance.

3.2 Defining prices

There have been tensions from the start between different objectives in opening the BnF's repository to other users, and it is no surprise that they resurfaced throughout the long process of setting the prices for archiving.

---

[9] "Association Française de Normalisation", that is, the French Association for Standardization.

[10] http://www.iso.org/iso/home/store/catalogue_tc/catalogue_detail.htm?csnumber=54911

[11] http://datasealofapproval.org.

First came the question of what the library wanted to sell: software, or a service? Around the time when SPAR was becoming operational, president Nicolas Sarkozy launch the idea of a "Great Loan", whereby the French State would borrow money to finance projects in new technologies, including the digital sector, and stave off economic crisis. As the Loan was shaping up to become the "Investissements d'avenir" (investing in the future) program in 2010, many public institutions were scrambling to set up projects that would fit what was known of the governmental action. As it was designed to boost the economy, sizeable returns on investment were expected, and the BnF sought out partners to monetize one of its big digital assets, SPAR, on a large scale. The potential partners who came forward thought of selling maintenance services around the software, to be used as a whole, or of making use of the BnF's vast digital storage facilities to bring down the costs of their services. Nothing came to fruition, although some talks are continuing along the same lines with other institutions. Yet it also meant that any ideas of making SPAR an open-source project, which would have required some initial spending and would not have yielded visible financial rewards, were put aside.

Meanwhile, the initial idea to have a track dedicated to third-party archiving within the BnF's instance of the system was not put in jeopardy by the discussions surrounding "Investissements d'avenir". It was, however, almost as difficult to price, first because the BnF, as a library, has little experience in selling goods and services, secondly, because the market for digital preservation services is still emerging, with very different offers, from cloud storage to comprehensive preservation, and not even private firms have a strong hold on their price range. The library decided to contract a consulting firm to get an idea of how much it could ask. The prices that emerged from the study then had to be validated by the ministries of Culture and Finance, who had their own priorities and policies.

Three factors were taken into account to set the prices for third-party archiving: existing pricing tiers on the market, willingness to pay, and the costs to the BnF of adding an extra terabyte of third-party data into the existing repository. Two price points were taken into account: direct costs due to the extra data (storage media, servers, manpower for ingest operations...), and global costs of maintaining the repository (software, hardware, expertise...), to be shared by the BnF and its customers.

Regarding the investment costs of preserving extra data, the consultants considered volumes, type of storage media (two tapes, two tapes and one disk, two tapes and two disk copies, and the benefits in terms of access gained with each extra disk copy), complexity of data ingest (from a self-serve dropbox to a tailored solution) and contract duration (a longer contract would level off the costs). The most recently acquired media were used as a basis to define the cost of the storage, brought back to the cost per Tb; a share of the costs of the tape libraries, tape readers and disk arrays was added. The human resources costs were assimilated to three days of an engineer's time should the client do most of the ingestion operation, ten to twenty days for a tailored solution. The costs of developing SPAR's software were calculated for a year, then divided by the number of existing terabytes.

As for the maintenance costs, they include support for the hardware, the software, the network and the sites, as well as a proportion of the human resources costs for daily operations, and assistance to the customer when needed (one and a half days for the generic ingest process, or three days for tailored solutions). On top of that is added 17.5% of the investments costs for one customer from the fourth year on, for maintaining the material acquired specifically.

Finally, as of Spring 2013, two tiers of clients have been identified:

- clients for the archiving services only;
- cultural institutions that have a partnership with the BnF as "Pôles associés" and benefit from other services (see 3.2.2).

3.2.1 Dedicated services: BnF Archivage numérique

Regular clients for the third-party archiving services will pay according to:

- the size of their collections, per terabyte. There is a decreasing price scale for 1Tb, then 2 to 5Tb, 6 to 9Tb, 10 to 29Tb and 30 to 49Tb;
- the number of copies they want made. The standard deal is for two copies on tape, one on each of the BnF's storage sites. One or two copies on disk come at an extra charge;
- the planned duration of archiving. So far, a decreasing price scale has been set for 3, 5 and 8 years;
- the level of service. Clients using the service autonomously, more or less as a drop box, pay less than those requesting evolutions in the code to have extra preservation functions. Those developments would in theory benefit the BnF's own collections as well, and so have been moderately priced[12].

3.2.2 Integrated services: Pôles associés

The missions of the Bibliothèque nationale de France include animating a national network of libraries[13]. As such, the BnF has distributed funding, first for catalog automation and integration to the national collective catalog, then for coordinated digitization programs. It seemed natural to promote preservation of these digitized collections. Members of the partnership programs will benefit from an 80% reduction in preservation costs if they entrust the BnF with the dissemination of their digitized materials in its digital library, Gallica. (Gallica already aggregates content from several institutions, whether through OAI-PMH indexation of content, or through the BnF's digitization programs, which include some digitized books and periodicals from partners.)

In addition, the BnF is building a Cooperation Portal extranet[14], to facilitate the management of different types of collaboration by the partners. A much-needed GUI for the monitoring of information packages' ingest, storage and dissemination is in the making, and could be a model for better communication between producers, preservation experts and repository administrators within the library as well.

3.3 Defining processes

PAIMAS[15] has been around for years (since 2004 as a CCSDS standard, 2006 as ISO 20652), and is still the only official, international standard for information exchange between the producers and the Archive. Yet the BnF has had trouble matching it to its own negotiation processes, mainly because of the many departments and teams involved in making the preservation

[12] http://www.bnf.fr/documents/archivage_num_tarifs.pdf.

[13] http://www.bnf.fr/en/professionals/national_cooperation/a.creating_national_network.html.

[14] http://espacecooperation.bnf.fr.

[15] Producer-Archive Interface Methodology Abstract Standard. Cf. http://public.ccsds.org/publications/archive/651x0m1.pdf.

services work. Potential clients are either sent to the Direction of Networks and Services if they are purely archiving clients, or to the Department of Cooperation if they are partners otherwise.

Preservation experts are in different departments according to their specialties. Different teams in the IT Department are involved when there are developments to be planned, on top of production planning to be sorted out. This is why the library has had to adapt PAIMAS to an idiosyncratic version, where phases can be aggregated, or distributed across several actors.

Meanwhile, as the BnF was contemplating courting the public archives community as clients, the Central Services for French Archives (Service Inter-ministériel des Archives de France, SIAF) published a standard for the exchange of data for archiving (Standard d'échange de données pour l'archivage, SEDA[16]). It has been developed since 2006, with version 1.0 published in September 2012, and a national standard is in the works with the name MEDONA. The standard describes formally the exchanges between the different actors during the archiving and retrieval of records, and provides an XML schema to encode the transactions. It has been created to facilitate the exchange of public records, in the realm of e-administration, between the services creating the information and the services in charge of archiving public data. Therefore its use is highly recommended to candidates seeking to sell short-term and mid-term preservation services of public archives. But the recent and rapid evolutions of the standard have led the BnF to put its implementation within the repository on hold, at least until the second semester of 2013.

4. CONCLUSION: WHAT'S NEXT?

Offers and prices for third-party archiving at the BnF are finally stabilized, with two tiers of clients, and this seems well positioned to benefit the library, through the incentive to develop new functionalities, and through some return on investment.

A first client, the Virtual Center of the National Museum of Modern Art, has led the way in taking up the offer, and this experience has helped streamline pricing, exchange processes, and workload management.

But how stable is the offer, really? The volatility of policies at the library and the state level carries an important risk at the management level. The existing clients' and partners' collections will be looked after according to contract, but what about the day-to-day operations' burden on the library's staff and resources? It is yet unclear whether the profits generated by the services will be enough to absorb the extra work, whether in setting up the administrative details of the contracts, dealing with the ingest and dissemination flows or adding new features to the repository, while maintaining an appropriate level of service for the BnF's own digital collections.

Moreover, the trend towards collaboration and sharing of resources is still being felt, as new projects emerge while budgets shrink. It is not clear at this stage which will prevail: the creation of multiple small repositories arising from the differences in size and constraints of the communities, even within the public sector, or the wish to regroup and save, and to share technology that is championed by its designers. Will the cost models be sustainable and guarantee the preservation of the partners' as well as the library's own collections?

Feedback from similar projects would help the management, as well as the team designing the software and the storage, assess its third-party archiving policy. Additional international benchmarking initiatives would benefit communities in a similar situation.

[16] http://www.archivesdefrance.culture.gouv.fr/seda/ (in French). The English presentation dates from the 2006, 0.1 version of the standard: http://www.archivesdefrance.culture.gouv.fr/seda/documentation/archives_echanges_v0-1_description_standard_v1-0-english.pdf.

A Risk Analysis of File Formats for Preservation Planning

Roman Graf
AIT - Austrian Institute of Technology GmbH
Donau-City-Strasse 1
Vienna, Austria
roman.graf@ait.ac.at

Sergiu Gordea
AIT - Austrian Institute of Technology GmbH
Donau-City-Strasse 1
Vienna, Austria
sergiu.gordea@ait.ac.at

ABSTRACT

This paper presents an approach for the automatic estimation of preservation risks for file formats. The main contribution of this work is the definition of risk factors with associated severity levels and their automatic computation. Our goal is to make use of a solid knowledge base automatically aggregated from linked open data repositories as the basis for a risk analysis in the digital preservation domain. This method is meant to facilitate decision making with regard to preservation of digital content in libraries and archives. We have developed a tool for aggregating rich and trusted file format descriptions. It exploits available linked data resources and uses expert models to infer knowledge regarding the long-term preservation of digital content. The ontology mapping technique is employed for collecting the information from the web of linked data and integrating it in a common representation. Furthermore, we employ AI techniques (i.e. expert rules, clustering) for inferring explicit knowledge on the nature and preservation-friendliness of the file formats. A statistical analysis of the aggregated information and the qualitative analysis of the aggregated knowledge are presented in the evaluation part of the paper. A Web service is created to support programmatic access to format and risk analysis reports.

Categories and Subject Descriptors

H.3.7 [**Digital Libraries**]: System issues; H.3.5 [**Online Information Services**]: Web-based services

Keywords

digital preservation, risk analysis, linked open data, preservation planning, ontology matching, information integration

1. INTRODUCTION

Preservation planning activities make use of the analysis and evaluation of file formats used for encoding digital content. The preservation risks for a particular file format are difficult to estimate and the definition of risk factors is still an open research topic. Intensive human expert involvement is required for searching and aggregating information about preservation risks and estimating of their possible impact in the future [2, 18]. The definition of risk factors for long term preservation can vary depending on preservation goals, workflows and assets used by a particular organisation. Also, the classification and weighting of risk factors is a challenging task, and is strongly dependent on the level of knowledge and experience of human experts. Individual domain specific knowledge bases do not contain all necessary semantic information required to perform an estimation of the preservation risks. The richness and the quality of knowledge base plays an important role in taking decisions on preservation planning. Even though the world wide web has turned out to be the largest knowledge base, the published information lacks a unified well-formed representation. The linked open data (LOD)[1] and Open Knowledge[2] initiatives address these weaknesses by defining guidelines for publishing structured data in standardized and queryable format. In order to aggregate sufficient knowledge about file formats for risk analysis we link together different independent and publicly available information sources like Freebase[3], DBPedia[4] and PRONOM[5].

The PRONOM registry provides persistent, unique, and unambiguous identifiers for file formats and therefore plays a fundamental role in the process of managing electronic records. Many file formats are properly documented, are open-source and well supported by producer. Other formats may be outdated, changed by software vendors and no longer functional with modern software or hardware. Some customized file formats could be obsolete and not accessible. To get a grip on all these problems we use the File Format Metadata Aggregator (FFMA) ([7]) system depicted in Figure 1, which aims at preparing the ground for knowledge base recommenders like DiPRec [6]. FFMA reuses the experience of building preservation planning tools and addresses the topic of digital long-term preservation. It performs an analysis of file formats based on the concept of risk scores. The knowledge base is built by following a linked data approach. Concretely, the information regarding file formats, software tools and vendors is retrieved from Freebase, DBPedia and PRONOM.

The important contribution of this paper consists in the technical information analysis and assessment regarding preservation risks for different formats. Another contribution is related to the usage of ontology mapping (see Figure 1) for the integration of different linked data sources into a common knowledge base. Decision support based on the elaborated rule engine provided by FFMA is meant to support institutions like libraries and archives with suggestions in the process of analyzing their digital assets. FFMA collects and structures information on file formats using a (semi-) au-

[1] http://linkeddata.org/
[2] http://www.okfn.org/
[3] http://www.freebase.com
[4] http://dbpedia.org/
[5] http://www.nationalarchives.gov.uk/PRONOM/

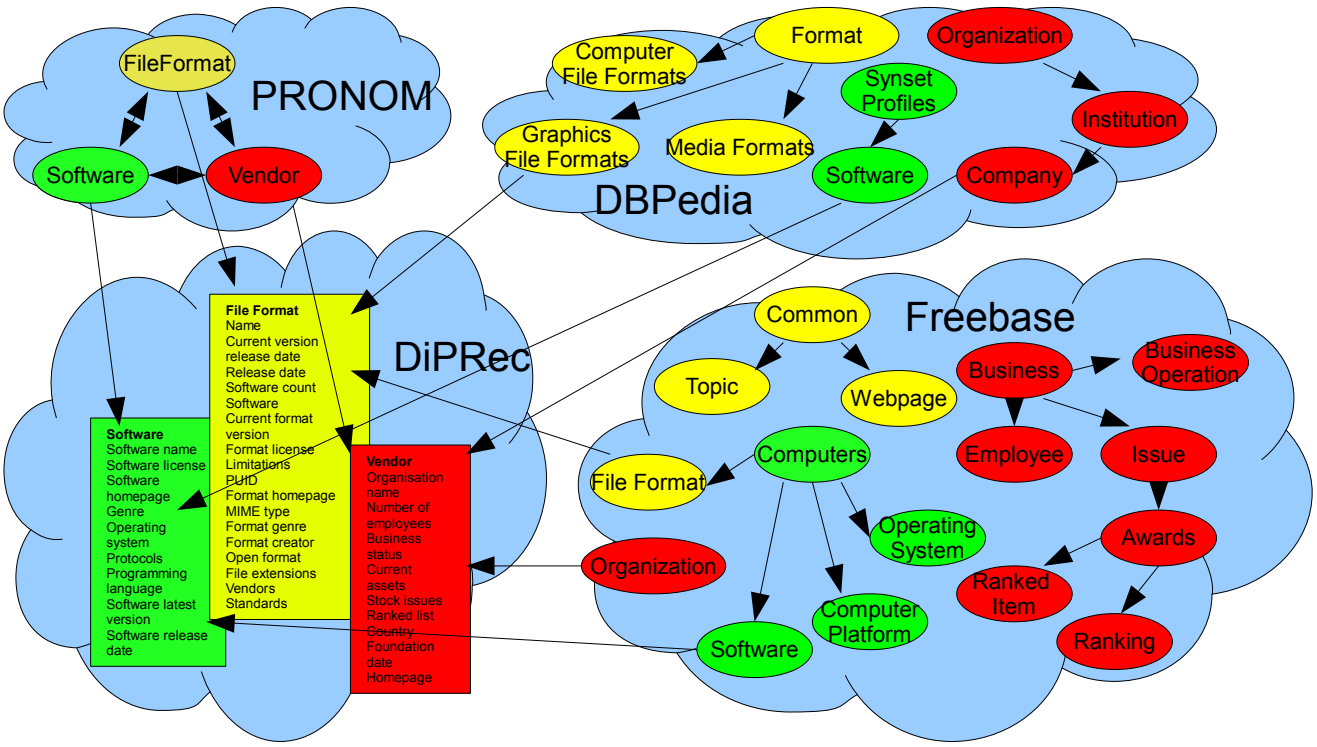


Figure 1: PRONOM, DBPedia and Freebase digital preservation domain related ontology sections mapped to the DiPRec file format ontology.

tomatic approach for knowledge extraction from the linked data repositories independent from the query language supported by individual repositories. We aim at designing well structured knowledge base with defined rules and scored metrics that is intended to provide decision making support for preservation experts. The paper is structured as follows: Section 2 gives an overview of related work and concepts. Section 3 explains knowledge base aggregation process and covers also ontology mapping, rule engine and algorithmic details of risk analysis. Section 4 presents the experimental setup, applied methods, description of the web service for risk analysis and results. Section 5 concludes the paper and gives an outlook about planned future work.

2. RELATED WORK

In [10] Andrew Jackson evaluated competing hypotheses regarding software obsolescence issue employing format identification tools for selecting appropriate preservation strategies. One of these hypothesis is presented by Rothenberg [17] and emphasizes that all formats should be considered brittle and transient, and that frequent preservation actions will be required in order to keep data usable. In contrast to that hypothesis the Rosenthal [16] claims that no one supporter of format migration strategy was able to identify even one format that has gone obsolete in the last two decades. Rosenthal argues that the network effects of data sharing inhibit obsolescence. But an accurate format identification and rendering is a challenging task due to malformed MIME types, rendering expenses, dependence on some content not embedded in the file, missing colour table, changed fonts, etc. In [10], the author examines how the network effects could stabilise formats against obsolescence in order to understand the warning signs, choices and costs involved. This evaluation should help to meet preservation strategy: either to perform frequent preservation actions to keep data usable or to concentrate on storing the content and using available rendering software. The result of evaluation demonstrates that most formats last much longer than five years, that network effects stabilise formats, and that new formats appear at a modest, manageable rate. However, he also found a number of formats and versions that are fading from use and that every corpus contains its own biases.

The PANIC tool [9] had the goal to automatically inform repository managers of changes that might cause risks for accessibility of their collections and alerting when file formats become obsolete. The idea of this tool was to aggregate data and metadata for further analysis, but this information is not easy browseable and the size of the repository is relative small in comparison to the LOD sources. Also there is no common understanding in the community about the meaning of term “obsolete” as mentioned above.

The AONS II tool [15] aimed at identifying file formats used for encoding digital collections, retrieving information regarding obsolescence risk indicators. The tool was building collection profiles and was referencing external format registries. This tool was able to distinguish accurately between

different versions of formats, in order to identify relevant risk levels. AONS II tool struggled to solve problems like misleading file extensions and different names for the same format by creating of internal format identifier for each apparent format found, and then tried to map it to the likely matching format identifiers used by external registries. But this tool did not apply risk factor metrics for risk calculation. Inspired by [15] we realized the need to develop a central web service that shares the results of local risk assessments with the community of interest. We aim at defining risk metrics based on experience of community members which share their individual expertise on defining and identifying risk factors. This would allow LOD registries to leverage the experiences and expertise of the contributing preservation community and add considerably to their usefulness.

The goal of the SPOT (Simple Property-Oriented Threat) model [18] is to identify previously unaddressed threats, perform preservation risk monitoring, and demonstrate the repository compliance to the accepted standards. In this work the digital preservation risks are divided into two categories: threats for preserving digital content, and threats for the custodial organization itself. The SPOT Model focuses on the first category and develops a framework for assessing threats arising from the technical operations associated with preserving digital objects. The SPOT risk model is limited to properties like availability, identity, persistence, renderability, understandability and authenticity. But these properties do not define measurable risk factors and do not exploit open knowledge from LOD repositories.

In the proposed approach we do not intend to mark down obsoleted formats, since there are different hypotheses and no common accepted definition for format obsolescence. Therefore we do not intend to treat obsolescence in a generalized form, but we treat it in an contextualized one. We define obsolescence in relation to the additional effort required to render a file beyond the capability of a regular PC setup in particular institution. This is consistent with the "institutional obsolescence" concept saying that a particular format that would not render anymore on a PC in an institution's reading room should be considered as obsolete. With FFMA we aim at assessing the risks associated with format rendering. We use the risk factors like "is compressed", "is supported by web browser", "has supporting software", "has supporting vendor", "is migration supported", "has digital rights information", etc. Most of these factors have influence on rendering the content. FFMA has the advantage of enabling users to configure the risk factors and scores according to their institutional context.

The format risk analysis approach in [5] presents the P2 registry, which is an RDF-based framework. The P2 registry employs information containing in DBPedia and PRONOM repositories and supports its own format risk analysis system. The main goal of the P2 platform is to allow and encourage publication of preservation data. This repository calls for the active participation of the digital preservation community to contribute data by simply publishing it openly on the Web as linked data. In contrast to the P2 registry the FFMA tool makes use of the rich Freebase repository as well and provides a modular architecture capable to easy integrate further repositories, even if they are not RDF based.

Additionally, FFMA uses a rule engine for risk analysis that handles further risk factors not covered by P2 registry and also supports their customization. Additional expert rules can be simply added to the model concept and the weighting severity levels are customizable as well.

Existing tools for long term preservation planning like Plato ([12, 1]) enable different digital preservation actions like identification, characterization and content migration. These tools present information about possible preservation action but do not provide suggestions or recommendations regarding format preservation risks for user that do not have an expertise in the digital preservation domain. The Plato tool defines decision criteria [3] for formats depending on institutional risk profile, but these criteria mainly are concentrated on format properties that can be obtained from P2 fact base and have predefined property values in contrast to normalized numerical values in FFMA expert system.

3. THE RISK ANALYSIS PROCESS

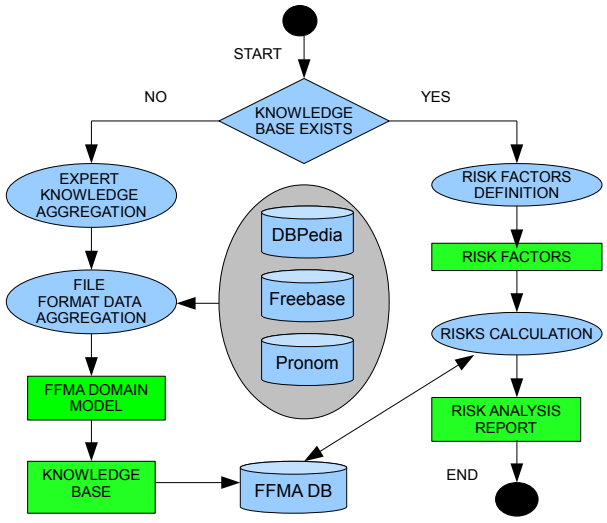


Figure 2: The format risk analysis workflow.

Figure 2 presents the risk analysis workflow. The building of the knowledge base (i.e. left side of the sketch) is a prerequisite for performing the risk computations. This includes the acquisition of expert knowledge and an aggregation of rich file format data. The creation of risk analysis reports is a two-step process based on the definition of risk factors and the computation and interpretation of risk scores (i.e. right side of the sketch). The result of risk calculation is presented in HTML format. The extended description of individual steps within this process is presented in the following sections.

3.1 Aggregation of File Format Data

The FFMA module for aggregation of file format descriptions collects information from LOD repositories and enhances it by using the expert knowledge aggregation module. At runtime, the aggregated metadata is processed and represented according to the underlying FFMA domain model

Aggregated File Format Description

DipFormat report for file format extension 'pdf'

DipId	Extension	PronomId	FreebaseId	DBPediaId	AitId
/Dip/pdf	pdf	fmt/14	/en/portable_document_format	<http://dbpedia.org/resource /Portable_Document_Format>	/Ait/Portable Document Format

Software

1. /Dip/ImageMagick 2. /Dip/Pdq Lite 3. /Dip/Adobe Acrobat 4. /Dip/Acrobat 5. /Dip/AutoCAD 6. /Dip/CorelDRAW 7. /Dip/FrameMaker 8. /Dip/Photoshop 9. /Dip/Illustrator 10. /Dip/Acrobat Messenger 11. /Dip/Acrobat Reader 12. /Dip/After Effects 13. /Dip/Framemaker 14. /Dip/GoLive 15. /Dip/InDesign 16. /Dip/Pagemaker 17. /Dip/Corel Ventura 18. /Dip/WordPerfect Office 19. /Dip/Photoshop Elements 20. /Dip/Crystal Info 21. /Dip/PDFedit 22. /Dip/ProTeXt 23. /Dip/KPDF 24. /Dip/Xpdf 25. /Dip/Beamer (LaTeX) 26. /Dip/Jfig 27. /Dip/PdfTeX 28. /Dip/LuaTeX

Vendors

1. /Dip/Adobe Systems 2. /Dip/<http://dbpedia.org/resource/Adobe_Systems>

Description

<pronom><fmt/14>Portable Document Format is a platform-independent format for representing formatted documents, developed by Adobe Systems Incorporated. It is the native format of Adobe's Acrobat family of software products, version 1.0 corresponding to the release of Acrobat 1.0. PDF is based on, and shares the same imaging model as, the PostScript page description language. A PDF file comprises a Header section, a Body section containing the objects which make up the document, a Cross Reference Table, and a Trailer section. PDF files can contain a wide variety of content, including text, images, video and audio.</fmt/14></pronom>
<dbpedia><<http://dbpedia.org/resource/Portable_Document_Format>>Portable Document Format (PDF) is a file format used to represent documents in a manner independent of application software, hardware, and operating systems. Each PDF file encapsulates a complete description of a fixed-layout flat document, including the text, fonts, graphics, and other information needed to display it. In 1991, Adobe Systems co-founder John Warnock outlined a system called " Camelot " that evolved into PDF. While the PDF specification has been available free of charge since at least 2001, PDF was originally a proprietary format controlled by Adobe. It was officially released as an open standard on July 1, 2008, and published by the International Organization for Standardization as ISO 32000-1:2008. In 2008, Adobe published a Public Patent License to ISO 32000-1 granting royalty-free rights for all patents owned by Adobe that are necessary to make, use, sell and distribute PDF compliant implementations.</<http://dbpedia.org/resource/Portable_Document_Format>></dbpedia>

Figure 3: Example of an aggregated data report for PDF file format.

by taking in account the configurations for a specific exploitation context. These configurations define which LOD repositories should be used and which file format properties are of interest for particular institutional context. The File Format Data Aggregation module is responsible for collecting descriptions on file format-related information from the open knowledge bases, while the FFMA engine combines the outcome of the module with the knowledge manually provided by domain experts after ontologies mapping in Expert Knowledge Aggregation module. The acquired domain knowledge in stored in a local database and further used in the reasoning risk computation process. We consider Freebase [13] as one of the most valuable sources for information extraction. It is a practical, scalable semantic database for structured knowledge. The PRONOM data format looks very similar to the FFMA ontology classes but it doesn't contain all necessary properties (like genre or vendor business status) that DiPRec requires to incorporate significant data from another ontologies. Extending the PRONOM repository, we collect information from additional sources and aggregate it in a homogeneous representation in the FFMA knowledge base, by using the FFMA domain model. The assignment to given property sets, the functions for value normalization, the queries for specific LOD repositories are the main constituent parts of the property definition model. An example of aggregated description for PDF format is presented in Figure 3. The external knowledge sources like DBPedia and Freebase manage huge amounts of LOD triples, which allows one to extract fragmental descriptions on file formats, software applications and software vendors. DBPedia allows to post sophisticated queries using SPARQL query and OWL ontology languages [11] for retrieving data available in Wikipedia. Public read/write access to Freebase is allowed through an graph-based query API using the Metaweb Query Language (MQL) [4]. PRONOM data is released as LOD and is accessible through a public SPARQL endpoint.

In order to reduce the required domain knowledge acquisition efforts the knowledge base stores the aggregated information in FFMA domain object model. After initial storage we only need to update specific database areas. This model increases performance because we do not need to perform expensive database queries with every operation. The potential drawbacks during the database initialization could be e.g. queries limit, bad internet connection to repositories or server could be offline for maintenance purposes. File format properties are designed to give an option at hand for

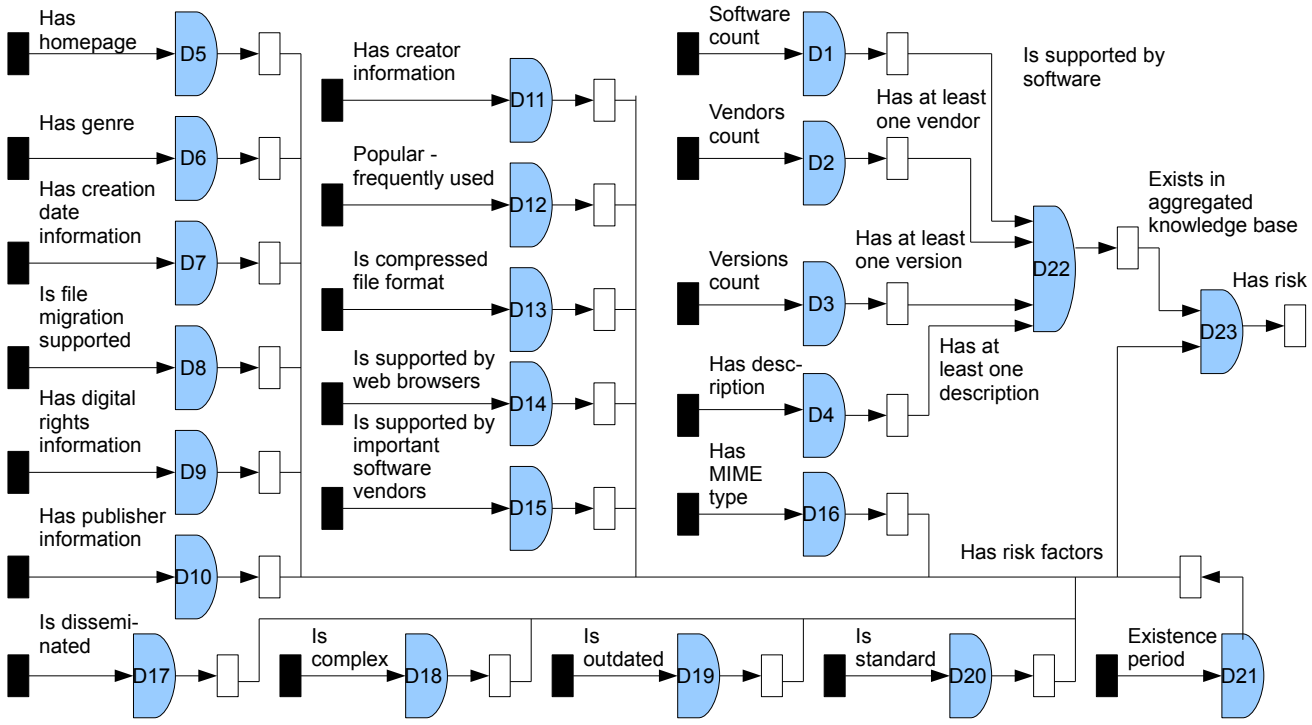


Figure 4: Forward rule chaining for risk analysis.

definition of user rules, metrics and classifications. The risk factors are used to compute overall preservation risks for a given file format.

3.2 Risk Factors Definition

The evaluation of the completeness of the knowledge aggregated from data registries (e.g. the percentage of file formats for which the genre property is available) gives some rough estimates about which risk properties should/can be defined and how to interpret them (i.e. weight and severity assignments). The most significant data repository queries in terms of digital preservation addresses PRONOM Id (i.e. PUID), file formats, software and software vendors. The properties of these main classes including computer platform, genre, license, programming language, release date, homepage, compression type and so on are of interest for risk analysis. Optionally, the user is able to extend the default risk analysis model by defining its own property sets of inferred knowledge and classifications using correspondent configuration files.

The information obtained from the digital preservation domain experts and from conducted experiments must be well structured. Typical scenarios were defined and the parameters used by library experts for collection handling were identified. Then linguistic labels were defined to classify measured values of each parameter and associated ranges. Finally, were determined the conditional rules that relate these linguistic labels to specific consequences. The knowledge acquisition for the Knowledge Base is performed by librarians who provide the knowledge engineer with typical application use cases, metrics and parameters that characterize the preservation processes [14] [8].

The most significant risk factors are related to the availability of software tools and vendors providing support for a particular file format (see Figure 4). For example, the version count metric could be interpreted in different ways. On the one hand the more versions a format has the more work is invested in its development and support. This implies that the given format is in use and well supported. On the other hand with the version count increases the probability that different versions will increase complexity and might generate conflicts when designing digital preservation workflows (e.g. for format migration purposes).

By changing severity values and classification settings, each customer could adjust the meaning of this risk factor for his specific context, needs and understanding. Documentation level is also an important risk factor. Additional help for risk estimation provide specification factors like whether a format has a homepage, genre definition, creator and publisher information, is supported by web browsers, has compression. The digital rights play increasingly important role in digital preservation. For preservation processes it is important to know whether format migration is supported. The MIME type provides a connection chain between different repositories. The complexity of the file format could be measured by assessment of documentation, format standard, relation between different versions of the same format, compression factor etc. Because the expert system contains information not only about format extensions but also about different versions, this knowledge could be covered by separate rule. Some formats are implicitly or explicitly declared as outdated or deprecated. The standardized formats have better chances of having a long time support. The time passed from the first release of a format is an additional metric for

File Format	Overall Risk Score	Overall Risk Level
pdf	0.14	Low

Detailed List of Format Risk Scores

Risk Factor	Property Value	Risk Score	Weight	Weighted Risk Score	Risk Level
Software Count	28	0.3	1.0	0.3	Middle
Vendors Count	2	0.0	1.0	0.0	Low
Versions Count	17	1.0	1.0	1.0	High
Has Description	2	0.3	1.0	0.3	Middle
Has MIME Type	true	0.0	0.2	0.0	Low
Format Existence Period	true	0.0	1.0	0.0	Low
Format is Complex	true	1.0	1.0	1.0	High
Format is Wide Disseminated	true	0.0	1.0	0.0	Low
Format is Outdated or Deprecated	false	0.0	1.0	0.0	Low
Has Genre	true	0.0	0.5	0.0	Low
Has Homepage	true	0.0	0.5	0.0	Low
Format is Open (standardised)	true	0.0	1.0	0.0	Low
Has Creation Date Information	true	0.0	1.0	0.0	Low
Is File Migration Supported	true	0.0	1.0	0.0	Low
Has Digital Rights Information	false	1.0	0.3	0.3	High
Has Publisher Information	true	0.0	0.1	0.0	Low
Has Creator Information	true	0.0	0.1	0.0	Low
Frequently Used (popular)	true	0.0	1.0	0.0	Low
Is Compressed File Format	false	0.0	0.9	0.0	Low
Is Supported By Web Browsers	true	0.0	0.5	0.0	Low
Is Supported By Important Software Vendors	true	0.0	0.3	0.0	Low

File Format	Overall Risk Score	Overall Risk Level
tif	0.26	Middle

Detailed List of Format Risk Scores

Risk Factor	Property Value	Risk Score	Weight	Weighted Risk Score	Risk Level
Software Count	135	0.0	1.0	0.0	Low
Vendors Count	1	0.3	1.0	0.3	Middle
Versions Count	9	1.0	1.0	1.0	High
Has Description	2	0.3	1.0	0.3	Middle
Has MIME Type	true	0.0	0.2	0.0	Low
Format Existence Period	true	0.0	1.0	0.0	Low
Format is Complex	true	1.0	1.0	1.0	High
Format is Wide Disseminated	true	0.0	1.0	0.0	Low
Format is Outdated or Deprecated	false	0.0	1.0	0.0	Low
Has Genre	true	0.0	0.5	0.0	Low
Has Homepage	false	1.0	0.5	0.5	High
Format is Open (standardised)	false	1.0	1.0	1.0	High
Has Creation Date Information	true	0.0	1.0	0.0	Low
Is File Migration Supported	true	0.0	1.0	0.0	Low
Has Digital Rights Information	false	1.0	0.3	0.3	High
Has Publisher Information	false	1.0	0.1	0.1	High
Has Creator Information	false	1.0	0.1	0.1	High
Frequently Used (popular)	true	0.0	1.0	0.0	Low
Is Compressed File Format	true	1.0	0.9	0.9	High
Is Supported By Web Browsers	true	0.0	0.5	0.0	Low
Is Supported By Important Software Vendors	true	0.0	0.3	0.0	Low

Figure 5: Sample risk reports for PDF and TIF file formats.

risk estimation. Mature and popular formats present lower preservation risk. Software, vendors and versions count factors together with a description factor build an aggregated rule whether the given format is supported by FFMA. Missing one of these important pieces of information means that the regarded LOD repositories do not provide information about required format.

The previously defined rules should be organized in order to process input statements (assertions) and to infer appropriate advice and conclusions. The forward rule chaining for file format analysis is presented in Figure 4. Forward chaining is the process of moving from the antecedents ("if" conditions) to the consequents ("then" actions) in a rule-based system. A specific rule is triggered if all of its inputs are available (i.e. a risk is present only if all assigned input properties are available). The antecedent is considered satisfied when the input values match the assertion, in which case the rule computes a risk value as consequent, otherwise a default risk value is set as consequent. Assertions are depicted by black rectangles on the input side and by the white rectangles on the output side (i.e. as result of the rule evaluation). The rules are presented by blue half-spheres, respectively. The output of one rule is used as an input for the following rule in the chain.

As an example, the rule-base system may start risk identification with the rule D1 supposing that software count is higher than 0. If the antecedent pattern defined in classification settings matches that assertion, the value x becomes "is supported by software" and the rule D1 fires. When the aggregated risk of rules D2, D3 and D4 matches the antecedent patterns for vendors, versions and descriptions count and has acceptable risk level severity, rule D22 fires, establishing that the format exists in aggregated knowledge base. This fact enables further analysis and similar iteration through remaining rules.

3.3 Risk Computation

The final conclusion of the rule-based system is whether an analysed file format has high, middle or low preservation risk and which particular risk factors cause this risk. The computation and interpretation of risk scores is completed within the Risk Calculation task (see Figure 2) by using the previously presented forward chaining model (see Figure 4). The risk score for a particular property is evaluated from risk analysis model dependent on metrics, property weight and risk interpretations. Each rule is responsible for the computation of a risk factor, and the weighted risk scores are used for computing the total risk score for a given format (see Figure 5 for an illustrative example).

Due to management and maintenance reasons, properties are grouped by sets. A property may belong to one or more property sets. The extent to which a property belongs to a property set and consequently contributes to the risk computation over a given dimension is modeled through the introduction of specific weighting factors (see Equation 1). The computation of the overall risk score for FFMA properties is presented in [6] and is computed as a weighted sum over all risk factors:

$$R_i = \sum_{ps \in PS_i} w_{ps,i} * \sum_{p \in PROP_{ps}} w_{p,ps} * d(p, PFV(p)) \quad (1)$$

Where R_i represents the preservation risk computed over the preservation dimension i, ps represents the index of the current property set within all sets associated to the dimension i (PS_i). The $w_{(ps,i)}$ is the weight of the contribution of the property set ps to dimension i. Similarly p stands for the index of current properties within the list of properties available in the given property set $PROP_{ps}$. $w_{p,ps}$ denotes the importance of a property p for the property set ps. The distance between the current property and the defined - 'preservation conform' - value for this property is represented through $d(p, PFV(p))$. The 'preservation con-

form' values and the metrics for distance computation are specified within the property definitions.

The final risk report contains detailed information about computed risk scores for each property, the weighting factors used in risk computations, the total risk scores for a file format and their user friendly interpretations (i.e. indication of severity levels). This kind of report provides a solid evaluation of the file format descriptions and estimates the preservation friendliness based on the interpretation of computed preservation risks.

4. EXPERIMENTAL EVALUATION

The evaluation of format risks was conducted with the FFMA knowledge base aggregated for development of DiPRec recommender. Our hypothesis is that file format data automatically aggregated from LOD repositories will provide the rule engine with valuable information and will enable risk estimation for different file formats. It is expected that the distribution of calculated format risk scores will match to the associated information that was found in the domain literature. The "low risk" marked formats should indicate the currently most reliable file formats for digital preservation workflows. One of the most important use cases for FFMA system is an evaluating of software solutions available for processing of the preservation plans and its assessment regarding preservation risk. A Web service was developed that automatically retrieves file format related data from LOD repositories and performs reasoning on collected information employing specified risk factors. The basis of this service relies on rich data descriptions retrieved from LOD repositories. The collected information is processed, normalized, integrated into the knowledge base of the service and subsequently classified in order to calculate risk scores for particular file format. The programming interface of this service supports querying for descriptions of the file formats, software, vendors and associated information. Service supports checking of availability of the information in the service database and retrieving data from LOD repositories if necessary. Service provides generation of rich format descriptions and a report on format risks.

4.1 Evaluation Data Set

For evaluation purposes a subset of 13 representative, well known file formats was selected. The *GIF*, *PNG*, *JPG*, *BMP* and *TIF* formats belong to the raster graphics genre. *MP3* is the most used audio format, while the *PDF* format is mostly used for document formats, having multiple versions and being well supported by Adobe Acrobat toolset. The *HTML* format also has multiple versions and is used for creation of Web pages. The *DOC* and *PPT* are Microsoft formats supporting creation of multimedia documents and presentations. Some outdated file formats are presented by *MAC*, *SXW* and *DXF*. The *MAC* is a bitmap graphic format for the Macintosh, one of the first painting programs for this OS, supporting greyscale-only graphics. The *SXW* is an outdated text format for OpenOffice, while *DXF* is a vector graphic format for AutoCAD.

Aggregated data reports are presented in HTML format by the FFMA service. An example for the PDF file format is presented in Figure 3. This report comprises the FFMA identifier */Dip/pdf*, the unique identifiers within external repositories describing the *pdf* format. According to the LOD principles, each linked data repository has its own mechanism for non-ambiguous referentiation of the managed entities represented by unique Web URLs. By having a reference in a correct format, a user is able to easily request the information from a web service. In this case, the PRONOM identifier is *fmt/14*[6], the Freebase one is */en/portable_document_format*[7] and DBPedia is *Portable_Document_Format*[8], respectively.

Additionally information about 28 different software tools and one vendor associated with this file format was aggregated and presented by their unique FFMA identifiers. Two LOD repositories provide different descriptions for the given file format. Since aggregated information is stored in a database, calculation time of the report demonstrates real-time performance (lower then a half of second on regular PCs). Aggregated reports on file formats contain information like "FileFormatDescription", "SoftwareName", "RepositoryName", "SoftwareHomepage", "SoftwareDescription" etc. FFMA returns evaluated software, vendor and risk report objects in HTML format. The processing of LOD objects supports storage, retrieval and analysis of information retrieved from Web repositories. This structured information is a knowledge base to be used for deriving preservation recommendations.

4.2 Computation of risk factors

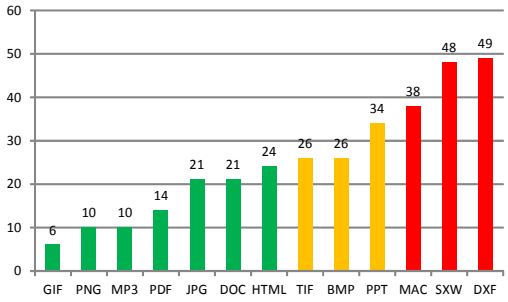


Figure 6: The distribution of the file formats with associated risk scores in range from 0 to 100 percent

Figure 6 demonstrates the distribution of the analyzed file formats according to their evaluated risk scores. The most reliable formats are marked by the green color, the middle risk formats with yellow color and the formats with the highest risks are flagged by the red color. Each format is also marked by its risk score in percent. In consequence, the experimental evaluation shows that *GIF* (6), *PNG* (10), *MP3* (10), *PDF* (14), *JPG* (21), *DOC* (21) and *HTML* (24) present the lowest preservation risks. The *TIF* (26), *BMP* (26) and PPT (34) formats have a middle preservation risk, while the *MAC* (38), *SXW* (48) and *DXF* (49)

[6] http://www.nationalarchives.gov.uk/PRONOM/Format/proFormatSearch.aspx?status=detailReport&id=613&strPageToDisplay=summary
[7] http://www.freebase.com/view/en/portable_document_format
[8] http://dbpedia.org/resource/Portable_Document_Format

Table 1: Exemplarily selected file formats with retrieved information for associated risk factors

Risk Factor	GIF	PNG	MP3	PDF	JPG	DOC	HTML	TIF	BMP	PPT	MAC	SXW	DXF
Software Count	18/M	21/M	12/M	28/M	17/M	164/L	39/L	135/L	18/M	4/M	122/L	1/H	9/M
Vendors Count	3/L	1/M	3/L	2/L	1/M	1/M	1/M	1/M	1/M	1/M	1/M	1/M	1/M
Versions Count	2/M	3/M	1/L	17/H	9/H	15/H	7/H	9/H	7/H	7/H	1/L	1/L	23/H
Has Description	2/M	2/M	1/H	2/M	1/H	2/M	1/H	2/M	1/H	1/H	1/H	1/H	1/H
Has MIME type	+/L	+/L	+/L	+/L	+/L	+/L	+/L	+/L	+/L	+/L	-/H	-/H	-/H
Existence Period	+/L	+/L	+/L	+/L	+/L	+/L	+/L	+/L	+/L	+/L	+/L	+/L	+/L
Is Complex Format	-/L	-/L	-/L	+/H	-/L	-/L	+/H	+/H	-/L	-/L	-/L	+/H	+/H
Is Wide Disseminated	+/L	+/L	+/L	+/L	+/L	+/L	+/L	+/L	+/L	+/L	-/H	-/H	-/H
Is Outdated or Deprecated	-/L	-/L	-/L	-/L	-/L	+/H	+/H	-/L	-/L	+/H	+/H	+/H	+/H
Has Genre	+/L	+/L	+/L	+/L	+/L	+/L	+/L	+/L	-/H	-/H	-/H	-/H	-/H
Has Homepage	+/L	-/H	-/H	+/L	-/H	-/H	-/H	-/H	+/L	-/H	-/H	-/H	-/H
Is Open (Standardised)	+/L	+/L	+/L	+/L	+/L	-/H	+/L	-/H	-/H	-/H	-/H	-/H	-/H
Has Creation Date	+/L	+/L	+/L	+/L	-/H	+/L	+/L	+/L	-/H	-/H	-/H	-/H	-/H
Has File Migration Support	+/L	+/L	+/L	+/L	+/L	+/L	+/L	+/L	+/L	+/L	+/L	+/L	+/L
Digital Rights Information	-/H	-/H	-/H	-/H	-/H	-/H	-/H	-/H	-/H	-/H	-/H	-/H	-/H
Has Publisher Information	+/L	-/H	+/L	+/L	+/L	+/L	+/L	-/H	+/L	-/H	-/H	-/H	-/H
Has Creator Information	+/L	-/H	+/L	+/L	+/L	+/L	+/L	-/H	+/L	-/H	-/H	-/H	-/H
Is Popular Format	+/L	+/L	+/L	+/L	+/L	+/L	+/L	+/L	+/L	+/L	-/H	-/H	-/H
Has Compression Support	-/L	-/L	-/L	-/L	-/L	-/L	-/L	+/H	-/L	-/L	-/L	-/L	-/L
Supported by Web Browser	+/L	+/L	+/L	+/L	+/L	+/L	+/L	+/L	+/L	+/L	+/L	+/L	+/L
Has Vendor Support	+/L	+/L	+/L	+/L	+/L	+/L	+/L	+/L	+/L	+/L	+/L	+/L	+/L
Total Risk (%)	6/L	10/L	10/L	14/L	21/L	21/L	24/L	26/M	26/M	34/M	38/H	48/H	49/H

formats were evaluated as being the less trusted ones. *BMP* and *TIF* have the same overall risk score by 26 percent, but this is a result of aggregating the weighted scores of different low-level risks. By breaking down the results in risk factors, one can verify that the *TIF* format has more descriptions, but in the same time it is more complex then the *BMP*. The genre information for *BMP* was not found (i.e. in the aggregated knowledge base), whereas for *TIF* no homepage link is available and a creation date was available only for *TIF*. In contrast to this, for *BMP* format the publisher and creator information is available. Additionally the *TIF* format is a compressed one, fact that increases its preservation risk.

The aggregated risk scores were computed by using the model described in Section 3 by employing the information aggregated within the knowledge base and computing individual risk factors relevant to the given file formats. Table 1 presents an overview of the computed low level risks for the formats included in the evaluation set. The values and the interpretations of the most important 23 risk factors are presented. Within this representation, the "+" sign stands for *true* while the "-" sign means *false*. *L* depicts low risk, *M* means middle risk and *H* stands for high risk. This table shows that among evaluated formats, the *DOC* format has the highest number of supported software, whereas for *SXW* only one software tool was documented in LOD repositories. The remaining formats have different software numbers, mostly between 10 and 40.

Therefore, the risk regarding the "software count" for *SXW* was considered as being high, the risks for *DOC*, *HTML*, *TIF* and *MAC* extensions as low and medium risk is associated with remaining formats. By defining classifications for this risk factor, it was expected that the more software tools support particular file format the lower is its risk. But this factor can be also configured according to the idea, that many software tools could cause instability of file format. In this case, the user may redefine classification settings according to his risk estimation preferences. The lowest risk for "vendors count" risk factor were calculated for *GIF*, *MP3* and *PDF* formats with two to three vendors. The remaining formats have middle value associated with this risk, in consequence no high risk regarding "vendors count" component was detected for the given data set. High vendor risk would be expected in the case that no vendors were documented for particular format. It was assumed that the more versions are defined for a format the higher is the probability of version confusion. Therefore our calculation evaluated the highest "versions count" factor risk for *DXF* (23), *PDF* (17), *DOC* (15), *JPG* (9), *TIF* (9), *HTML* (7), *BMP* (7) and *PPT* (7). Regarding availability of textual descriptions, it was expected that the more information was found, the lower is the risk. According to this risk definition the high risks were detected for *MP3* (1), *JPG* (1), *HTML* (1), *BMP* (1), *PPT* (1), *MAC* (1), *SXW* (1), *DXF* (1) formats and middle description factor risk with values in range from two to three for remaining formats. All of the regarded formats have multiple descriptions but do not exceed threshold of three and therefore there is no low risk among them. The MIME type is an essential reference in order to address a file format and to create a connection between different file format ontologies or identification tools. Most of the presented formats have found an associated reference. Only three formats are missing the MIME type: the *MAC*, *SXW* and *DXF* formats. The longevity of the format existence period could give us a rough estimation about its stability. Therefore the longer a format is in use the lower is the preservation risk. In our case all of the formats have low risk in this regard. The complexity of the format could cause additional preservation risks. Complexity here means the compatibility between different format versions, semantic information necessary for correct rendering, using of compression, missing standard or documentation. In our list as complex formats were marked *PDF*, *HTML*, *TIF*, *SXW* and *DXF*. The dissemination level plays an important role in development of associated software tools and popularity of the format. In this regard high preservation risk have *MAC*, *SXW* and *DXF*. Some formats in the associated literature and in expert community are marked as outdated or deprecated due to limited using of this format or some of its versions. These formats are *DOC*, *HTML*, *PPT*, *MAC*, *SXW* and *DXF*. The open or standardised formats have lower preservation risks like *GIF*, *PNG*, *MP3*, *PDF*, *JPG* and *HTML*. Formats that have homepage have lower risks due to additional information placed in their homepages. Our tool found homepages for three formats *PDF*, *GIF* and *BMP*. These formats therefore are regarded as having lower risks. The genre information also reduce risks for *GIF*, *PNG*, *MP3*, PDF, *JPG*,

DOC, *HTML* and *TIF*. The creation date factor could be implemented in different ways. In our meaning the older is the file format the more it was used and the more stable it is. Therefore *GIF*, *PNG*, *MP3*, *PDF*, *DOC*, *HTML* and *TIF* have low risk expectation in this regard. Other researchers could consider the latest date as more reliable. Another important aspect for digital preservation is an ability to migrate file from one format to another. In this regard all of examined files have low risk in regular institutional environment. Digital rights information plays increasingly important role in digital preservation. Extraction of this important information is a topic of future work. Publisher and creator information gives us additional source to decide how much trust should be given to the particular publisher. Our risk analysis tool found the information required for *MP3*, *DOC*, *HTML*, *PDF*, *GIF*, *BMP* and *JPG*. In order to evaluate how frequently particular format is used in libraries preservation workflows was used expert knowledge. The most popular formats are *GIF*, *PNG*, *MP3*, *PDF*, *JPG*, *DOC*, *HTML*, *TIF*, *BMP* and *PPT*. In order to accumulate expert knowledge like in case of frequently used formats was designed new data repository that provides information missed in other LOD repositories. Similarly the compression support, web browser support and vendor support information is a topic of future work.

The different risk scores for *DOC* (low) and *PPT* (middle) could be explained with larger amount on software tools automatically detected for *DOC* (164) comparing to four for *PPT* and also with more descriptions for *DOC* format. Additionally, for *DOC* the genre, creation date, publisher and creator information were retrieved, whereas these factors are missing for *PPT*. This does not mean that such information does not exist for PPT, it only indicates that this is not included or not found in LOD repositories. The same consideration is valid for the "software count" value 12 of *MP3* format. It is known that there should be much more associated software tools that are able to handle this format.

At this point it should be stated that not all formats were analyzed and that evaluated results currently require verification by human expert and further optimisation of calculation methods. Evaluation results presented in Figure 6 and Table 1 are limited to the information automatically collected from mentioned above LOD repositories and is customized by applied expert rules. Therefore these results cannot be regarded as absolutely accurate, but they provide a good overview of the possible preservation risks related to the given file formats. The classification settings for risk factors are institutional dependent and is a matter of discussion and a future work. The default thresholds are defined based on the accessible expert knowledge and could be customized according to preferences of particular user.

4.3 Web service for risk analysis report

Finally, the presented approach was implemented as a REST-Full web service, allowing individuals and third party applications to make use of available risk computations[9]. We aim also at collecting more user feedback and to improve the presented risk computation models. Figure 5 presents user friendly presentations of the analysis reports regarding the *PDF* and *TIF* file formats. The *PDF* format has the low preservation risk with 14% and the *TIF* format has the middle preservation risk with 26%. The report includes the nominal values for the risk properties, their weighting in risk computations, the derived risks scores, the individual interpretations (i.e. risk level) and their weighting for the computation of the total risk score. In the provided examples, the most significant risk factors like software count, vendors count, versions count, standardisation, popularity, description factor, creation date factor and migration factor have the highest weight 1.0; the less important factors have weights in range between 0.1 and 0.5. The risk analysis reports provided by Web service demonstrate that our hypothesis was correct. The file format descriptions automatically aggregated from LOD repositories provide sufficient information to enable estimation of preservation risks for various file formats. The distribution of calculated format risk scores proves that file formats flagged as "low risk" formats are (still) reliable file formats. Old, outdated formats like *SXW* or *DXF* were identified as presenting increased preservation risks by the given models.

[9]`http://ffma.ait.ac.at:8080/preservation-riskmanagement/`

5. CONCLUSIONS

This paper presents the risk analysis service which employs FFMA knowledge base with rich descriptions of computer file formats. The service uses semi-automatic information extraction from the Linked Open Data repositories, analyzes and aggregates knowledge that facilitates decision making in different institutions for preservation planning. The main contribution of this paper is the definition of the risk factors, their automatic computation and interpretation based on aggregated knowledge base. The FFMA knowledge base is created using the ontology mapping approach for collecting data from LOD repositories. This allows automatic retrieval of rich, up-to-date information, reducing the setup and maintenance costs for the risk analysis service. Since the knowledge acquisition and aggregation process is automated, this will allow the aggregated knowledge base to be easily updated. The scalability of information extraction was improved by reducing the domain knowledge acquisition efforts by means of storing the aggregated knowledge in a local database. The evaluation of the preservation friendliness is based on the expert models employed for performing the computation of risk scores. The underlying expert model can be easily adapted to the preservation requirements of particular institutional contexts through the customization of the configuration files, the risk definitions and their associated severity levels. A Web service was implemented to support the evaluation of the aggregated knowledge base and to support decision making on digital preservation actions based on the provided risk analysis reports. The evaluation part of the paper presets the computation of risk analysis reports for a representative set of 13 well known file formats. The presented model makes use of 23 different risk factors. The interpretation of experimental results demonstrates the viability of the proposed approach. Anyway, there are still two main drawbacks of the proposed approach. The first of them is related to the need to reason based on incomplete information (e.g. the description of file formats is not complete in either of the given repositories). The second one is related to the need to adjust the weighting of the risk factors according to individual institutional contexts.

As future work we plan using of additional knowledge sources (e.g. vendor's web sites, further knowledge bases) and additional properties for format descriptions (e.g. popularity of file formats available on `http://www.fileinfo.com/`). The extension of expert rules with new risk factors, improving the accuracy of the expert model and enhanced identification of software tools supporting individual file formats are additional research topics to be investigated.

6. ACKNOWLEDGMENTS

This work was supported in part by the EU FP7 Project SCAPE (GA#270137) `www.scape-project.eu` and partially by the EU project "ASSETS - Advanced Search Services and Enhanced Technological Solutions for the European Digital Library" (CIP-ICT PSP-2009-3, Grant Agreement n. 250527). The authors wish to thank Paul Wheatley from the British Library for his thoughts on the topic.

7. REFERENCES

[1] B. Aitken, P. Helwig, A. Jackson, A. Lindley, E. Nicchiarelli, and S. Ross. The planets testbed: Science for digital preservation. *Code4Lib*, 1(3), 2008.

[2] P. Ayris, R. Davies, R. McLeod, R. Miao, H. Shenton, and P. Wheatley. The life2 final project report. Final project report, LIFE Project, London, UK, 2008.

[3] C. Becker and A. Rauber. Decision criteria in digital preservation: What to measure and how. *Journal of the American Society for Information Science and Technology (JASIST)*, 62(4):1009–1028, 2011.

[4] C. Bizer, T. Heath, and T. Berners-Lee. Linked data - the story so far. *Int. J. Semantic Web Inf. Syst.*, 5(3):1–22, 2009.

[5] L. C. David Tarrant, Steve Hitchcock. Where the semantic web and web 2.0 meet format risk management: P2 registry. *International Journal of Digital Curation*, 6(1):165–182, 2011.

[6] S. Gordea, A. Lindley, and R. Graf. Computing recommendations for long term data accessibility basing on open knowledge and linked data. *Joint proceedings of the RecSys 2011 Workshop on Human Decision Making in Recommender Systems (Decisions@RecSys'11) and User-Centric Evaluation of Recommender Systems and Their Interfaces-2 (UCERSTI 2) affiliated with the 5th ACM Conference on Recommender Systems*, 811:51–58, November 2011.

[7] R. Graf and S. Gordea. Aggregating a knowledge base of file formats from linked open data. *Proceedings of the 9th International Conference on Preservation of Digital Objects*, poster:292–293, October 2012.

[8] S. Hoorens, J. Rothenberg, C. van Oranje, M. van der Mandele, and R. Levitt. Addressing the uncertain future of preserving the past: Towards a robust strategy for digital archiving and preservation. Technical report, RAND Corporation, 2007.

[9] J. Hunter and S. Choudhury. Panic: an integrated approach to the preservation of composite digital objects using semantic web services. *International Journal on Digital Libraries*, 6, (2):174–183, September 2006.

[10] A. N. Jackson. Formats over time: Exploring uk web history. *Proceedings of the 9th International Conference on Preservation of Digital Objects*, pages 155–158, October 2012.

[11] L. Jens, S. Jörg, and A. Sören. Discovering unknown connections -the dbpedia relationship finder. In *Proceedings of the 1st Conference on Social Semantic Web (CSSW)*, volume P-113, pages 99–109, Leipzig, Germany, 2007. Gesellschaft für Informatik.

[12] R. King, R. Schmidt, A. Jackson, C. Wilson, and F. Steeg. The planets interoperability framework: An infrastructure for digital preservation actions. In *ECDL09 Proceedings of the 13th European conference on Research and advanced technology for digital libraries*, volume 5714/2009, pages 425–428. Springer-Verlag, 2009.

[13] B. Kurt, E. Colin, P. Praveen, S. Tim, and T. Jamie. Freebase: a collaboratively created graph database for structuring human knowledge. In *SIGMOD '08 Proceedings of the 2008 ACM SIGMOD international conference on Management of data*, pages 1247–1249, New York, NY, USA, 2008. ACM.

[14] A. McHugh, S. Ross, P. Innocenti, R. Ruusalepp, and H. Hofman. Bringing self-assessment home: Repository profiling and key lines of enquiry within drambora. *International Journal of Digital Curation*, 3(2), 2008.

[15] D. Pearson and C. Webb. Defining file format obsolescence: A risky journey. *The International Journal of Digital Curation*, Vol 3, No 1:89–106, July 2008.

[16] D. S. Rosenthal. Format obsolescence: assessing the threat and the defenses. *Library Hi Tech*, 28(2):195–210, 2010.

[17] J. Rothenberg. Digital preservation in perspective: How far have we come, and what's next? *Future Perfect 2012*, 2012.

[18] S. Vermaaten, B. Lavoie, and P. Caplan. Identifying threats to successful digital preservation: the spot model rsik assessment. *D-Lib Magazine*, 18(9/10), September 2012.

On the Assessment of Preservability: Method and Application

Diogo Proença, Gonçalo Antunes
IST/INESC-ID
Lisbon, Portugal
{diogo.proenca, goncalo.antunes}@ist.utl.pt

Tomasz Miksa
Secure Business Austria
Vienna, Austria
tmiksa@securityresearch.at

ABSTRACT
This paper aims to establish engineering processes and methods for the assessment and deployment of digitally preservable systems by identifying a method for assessing the preservability capabilities of systems. The work done on this was based on the hypothesis that preservability consists of a set of systems capabilities that originates from a combination of system/software capabilities as defined in ISO 25010:2010. Based on that hypothesis, it was verified that such quality characteristics influence the preservability of systems. That influence is relative, since it depends on the specific scenario being addressed and on the concerns and requirements of the stakeholders of the system, as different qualities of a system might assume different degrees of importance along time. With those principles taken into account, this work developed an assessment method for assessing the preservability of systems that can be adapted to each scenario being analyzed. For demonstrating the application of the method, an example assessment was performed on a specific scenario, which resulted on the revelation that preservability of the system in focus on that particular case can be greatly improved.

Categories and Subject Descriptors
H.1 [**Information Systems**]: Models and Principles; J.1 Administrative Data Processing Government; K.6.4 Management of computing and Information Systems.

General Terms
Management, Documentation, Measurement, Verification.

Keywords
Trust, Digital Preservation, Checklist Assessment.

1. INTRODUCTION
It can be said that the successful preservation of a business process depends on the capturing of context that is sufficient to be able to redeploy it in the future. However, different scenarios present different challenges considering the availability of that context for preservation. The technological context is such an example: some systems possess the necessary capabilities that make their preservation possible while others do not. In order to be able to easier distinguish between these systems, we introduce the concept of preservability.

We define **preservability** as the degree to which a system, product, or component can be archived for as long as necessary, ensuring its trustworthiness, and redeployed and re-executed according to the expectations, in a future environment, that might potentially be different from the original. This definition hints at the fact that the degree of preservability is always dependent on the requirements of the stakeholders, or in other words, it is dependent on the specific scenarios approached: different scenarios have different stakeholders with different needs concerning preservation. For instance, in some scenarios stakeholder‘s requirements might dictate that full functionality has to be preserved, while in other scenarios partial functionality might suffice.

Based on this definition, one can say that preservability seems to be a desired quality of systems, since it is usually not imposed by functional or business requirements. In fact, the hypothesis raised by this work claims that preservability is a set of system capabilities originating from a combination of system/software qualities. These system/software qualities assume different relevance among them depending on the scenario being assessed.

The subject of system/software qualities has decades of research. One of the most relevant references on this subject is the *ISO 25010:2010 – Systems and Software Engineering -- Systems and Software Quality Requirements and Evaluation (SQuaRE) – Systems and Software Quality Models10 [1]*. The standard defines a set of quality models that can be used in the identification of relevant system/software quality characteristics that can be further used to establish requirements, criteria for satisfaction and measures.

Based on the hypothesis raised, this work aims to define a method for the identification and assessment of relevant software qualities from the perspective of preservability. For that, the ISO qualities will be analyzed from the point of view of preservability. An assessment process based on *ISO 15504 - Information technology — Process assessment [2]* will be presented. Finally, a Civil Engineering Institution will be used to depict how the preservability of a real system could be assessed.

2. THE QUALITY CHARACTERISTICS OF SYSTEM’S PRESERVABILITY
The hypothesis raised by this paper is that preservability is attained via a set of system’s capabilities achieved by a combination of quality characteristics of systems. However, the assessment of preservability itself can be seen as a hard task since it involves present verification of something what can only be assured with full certainty in the future. Nonetheless, in order to be better prepared for being preserved and later redeployed in the future, systems should possess determined qualities.

The ISO 25010 defines quality characteristics for software systems which can be “further used to establish requirements, their criteria for satisfaction and the corresponding measures” [1]. It defines eight system/software product qualities: functional suitability, performance efficiency, compatibility, usability,

reliability, security, maintainability, and portability. In this section we relate these qualities to preservability, the definitions can be found in [1].

2.1 Functional Suitability

Functional suitability in terms of preservability, in some scenarios assumes particular importance, since the stakeholders might require that the system is fully functional when redeployed. The following aspects are considered sub-characteristics of functional suitability according to the ISO:

Functional completeness: In the perspective of preservability, this characteristic might assume particular importance in specific scenarios, where stakeholders require a fully functional redeployed system. In other scenarios, full functional completeness might not be so important, since the stakeholders might require only partial functionality to be redeployed.

Functional correctness: Concerning preservability, this characteristic might influence the decision to preserve. For instance, if a high degree of correctness is required by stakeholders, and if the system is not able to comply with it, then its preservation of the system might be ruled out.

Functional appropriateness: In terms of preservability, if the system does not possess this characteristic, then its stakeholders might not consider it particularly fit to be preserved.

2.2 Performance Efficiency

Performance efficiency in some scenarios this characteristic assumes particular importance, especially in scenarios where the stakeholders expect that the experience with the system remains unchanged. The following aspects are considered sub-characteristics of performance efficiency according to the ISO:

Time behavior: Concerning preservability, this characteristic becomes crucial if the stakeholders require the system's response and processing times to remain the same.

Resource utilization: Concerning preservability, the resource utilization might impact the choice to preserve or not a system, due to the high or low amounts of resources required or the expected availability of some types of resources in the future.

Capacity: In terms of preservability, this characteristic might assume importance in some scenarios since a system with greater capacity, might require more resources at the time of preservation, while a system with lower capacity might require less resources.

2.3 Compatibility

Compatibility is a very important aspect of digital preservation as after redeployment there is the need to assure that a system will perform as expected, despite having differences in the environment. Any incompatibility with other systems or any external dependency will endanger the preservability status of a system as the system might not perform as it was expected. The following aspects are considered sub-characteristics of compatibility according to the ISO:

Co-existence: In terms of preservability this attribute can be used in conjunction with the dependency capturing to check for possible dependencies that are critical for the correct execution of the system. It will also help to check if there are any incompatibilities between the system and other products, so that in the future we can use all this data to guarantee the correct execution of the system and eliminate the existence of any incompatibility.

Interoperability: In terms of preservability this attribute can be used to assess to what extent a certain system makes use of proprietary protocols, which can endanger the preservability of the system, due to licensing or third-party systems needed. This attribute can also be used to check with which other systems our system is communicating with and are essential for the correct execution of it, making it useful for dependencies capturing. Finally this attribute, can also be used as a measure of good communication channels between the system and other components which can enhance its preservability status.

2.4 Usability

Usability concerning preservability, usability might be important in determined scenarios with systems that involve heavy user interaction. The following aspects are considered sub-characteristics of usability according to the ISO:

Appropriateness recognisability: Depending on the scenario, this characteristic might impact the choice of doing preservation if, for instance, the scenario at hand requires or not the system to be appropriate for its users. This characteristic might also impact the success of adoption by future users after redeployment.

Learnability: Depending on the scenario at hand, it might impact the decision to preserve and might also impact the adoption by users after redeployment.

Operability: Depending on the scenario at hand, this characteristic might impact the success of adoption by future users after redeployment.

User error protection: Depending on the scenario at hand, this characteristic might impact the success of adoption by future users after redeployment.

User interface aesthetics: Depending on the scenario at hand, this characteristic might impact the success of adoption by future users after redeployment.

Accessibility: Depending on the scenario at hand, this characteristic might impact the success of adoption by future users after redeployment: if a system is currently difficult to use by a wide range of users, then it is probable that it will remain like that after redeployment.

2.5 Reliability

Reliability concerning preservability, might influence the decision to preserve a system. The following aspects are considered sub-characteristics of reliability according to the ISO:

Maturity: Concerning preservability, this characteristic might influence the decision to preserve a system, in the sense that a more mature system, will lead to less complications when preserving and redeploying.

Availability: Concerning preservability, this characteristic might influence the decision to preserve a system in certain scenarios, since a system which shows poor availability rates might not be considered for preservation.

Fault tolerance: Concerning preservability, this characteristic might influence the decision to preserve a system in certain scenarios, since a system which shows poor fault tolerance rates might not be considered for preservation.

Recoverability: This characteristic might be very important for preservability since it might facilitate the preservation and redeployment of the system.

2.6 Security

Security is a crucial aspect of digital preservation itself, since its impact might be positive or negative on the preservability of systems. Systems might manage sensitive data that should be considered when doing preservation. Additionally, systems might include mechanisms that can become troublesome to preservation. The following aspects are considered sub-characteristics of security according to the ISO:

Confidentiality: Confidentially might impact negatively the preservability of a system if, for instance, encryption mechanisms are being used in the system for securing accesses to the data, which would also involve preserving the encryption keys. Additionally, confidentiality might involve the use of external systems for managing the access to files.

Integrity: Integrity is considered a basic property of DP. In terms of preservability, it is desirable that a system has built-in integrity mechanisms, since that can be a guarantee that either the system or the data have not been changed in an unauthorized way prior to preservation. Integrity should then be ensured during the archive phase.

Non-repudiation: In terms of preservability, it is desirable that non-repudiation is ensured when preserving a system, since the historic of all actions or events happening before the system was preserved is important to ensure the provenance of the system and its data, ensuring the authenticity of the preserved objects. Provenance is necessary to validate the authenticity of preserved data, and includes the documented history of creation, ownership, accesses, and changes occurred over time.

Accountability: In terms of preservability, it is desirable that accountability be ensured when preserving a system, since the historic of all actions or events happening before the system was preserved, and its relation with different entities concerned with the system, is important to ensure the authenticity of the system and its data.

Authenticity: In terms of preservability, authenticity concerns the reliability of the objects in the sense that the control over their custody is enforced [1]. As such, it is a basic property of DP and often includes the existence of mechanisms for authentication and authorization as a way of enforcing it.

2.7 Maintainability

Maintainability is the ease of reconfiguration of the running system, product, or component by its maintainers and ability to cope with a changed environment. In case of digital preservation, the maintainers are the persons responsible for redeployment and the changed environment is the redeployment environment. When a system, product, or component is being redeployed it has to be fitted into the existing environment. The possibility to influence several settings and parameters of a system, product, or component increases the chance to redeploy it successfully. For example if a software which uses a database and external services have locations and addresses not hardcoded and therefore possible to modify, then the software has higher maintainability and higher preservability. However, in order to be able to benefit from maintainability a sufficient set of information describing the potential changes must be documented and preserved. Otherwise, high maintainability may decrease the preservability. The following aspects are considered sub-characteristics of maintainability according to the ISO:

Modularity: A System, product, or component with high modularity allows easy distinguishing between the modules. When any problems during redeployment occur, it is easier to deal with them within a module ("divide and conquer") rather than trace and identify their effects in the whole complex system, product, or component. Furthermore, different digital preservation actions may be suitable for different kinds of modules. Higher customization of digital preservation actions stemming from modularity should result in higher preservability.

Reusability: The higher reusability of a system, product, or component, the higher the likelihood that it is already a part of a knowledge base or a repository and therefore does not have to be the subject of digital preservation actions. Reusability of a system, product, or component may benefit from higher standardization. Furthermore, reusable systems, products, or components usually have broader community of users and thus more know-how and experience in preservation of these systems, products and components is available.

Analyzability: This is one of the critical requirements for preservability. The more information on execution of a system, product, or component is provided, the better the preservation actions can be adjusted. For example high analyzability facilitates identification of modules and their dependencies. Moreover, higher analyzability fosters the verification of system, product, or component redeployment. If some of modules cannot be redeployed, it may provide essential information to locate or reengineer substitute modules. Mechanisms like tracing, logging or provenance collection increase analyzability.

Modifiability: Modifiability is highly coupled with modularity and analyzability. It is very likely that, the more modular and analyzable the system, product, or component is, the easier it is to introduce and evaluate the modification. A need to modify the system, product, or component may occur when the preserved system, product, or component must be adjusted to the new redeployment environment, e.g. the database engine has to be substituted with a different one available on a different address.

Testability: While high analyzability allows passive collection of information, high testability allows active examination of a system, product, or component without affecting its state. It gives a possibility to design and run tests in the original environment. These tests can be executed in the redeployment environment and its results can be compared against original one. Moreover, testability allows verifying if any of introduced changes, like component substitution, are not affecting the system, product, or component in an undesired way.

2.8 Portability

Portability is one of the key aspects of digital preservation. In order to preserve a system we have to archive it and later redeploy it in a different environment. In this sense, a system with a high degree of portability will be highly desirable as it can be transferred without major incompatibilities of hardware, software or environment which will enhance its preservability status. The following aspects are considered sub-characteristics of portability according to the ISO:

Adaptability: In terms of preservability this attribute can be used to assess to what extent a system is prepared for different software, hardware or environments that might appear in the future. This is also a measure that can guarantee platform and hardware independence.

Installability: In terms of preservability this attribute can be important as a measure of easiness of installation of a certain system. An easy to install system is desirable as it ensures that there is fewer or no need for trained personnel to install the system, and reduces the total redeployment time. Moreover, if an installation procedure exists where is described how to install the system and/or automated installer exists it will enhance a system's preservability.

Replaceability: In terms of preservability this attribute can be used for alternatives assessment. In case other system or part of a system fails we can replace that system with an identified replacement system that will perform in the same way of the failing system. During redeployment, in case we can't redeploy the system due to missing dependencies, or any other reason, we can redeploy or use another system which was previously identified as replacement.

3. ASSESSMENT METHOD

This section contains the assessment process based on the guidance provided by ISO15504 [2], and serves as guidance on the nature of process required to assess preservability. The content of this process contains the minimum elements of a documented assessment process applicable for use in the context of assessing preservability.

Although this process includes only the activities, their description implicitly contains the other elements that may comprise a process: purpose, initial conditions, end condition, inputs, outputs, and roles and responsibilities.

The assessment process consists of the following activities: (1) Initiation, (2) Planning, (3) Briefing, (4) Data collection, (5) Data validation, (6) Analysis of the Preservability Assessment, and (7) Assessment reporting.

These activities are combined to form the assessment process for preservability depicted in Figure 1.

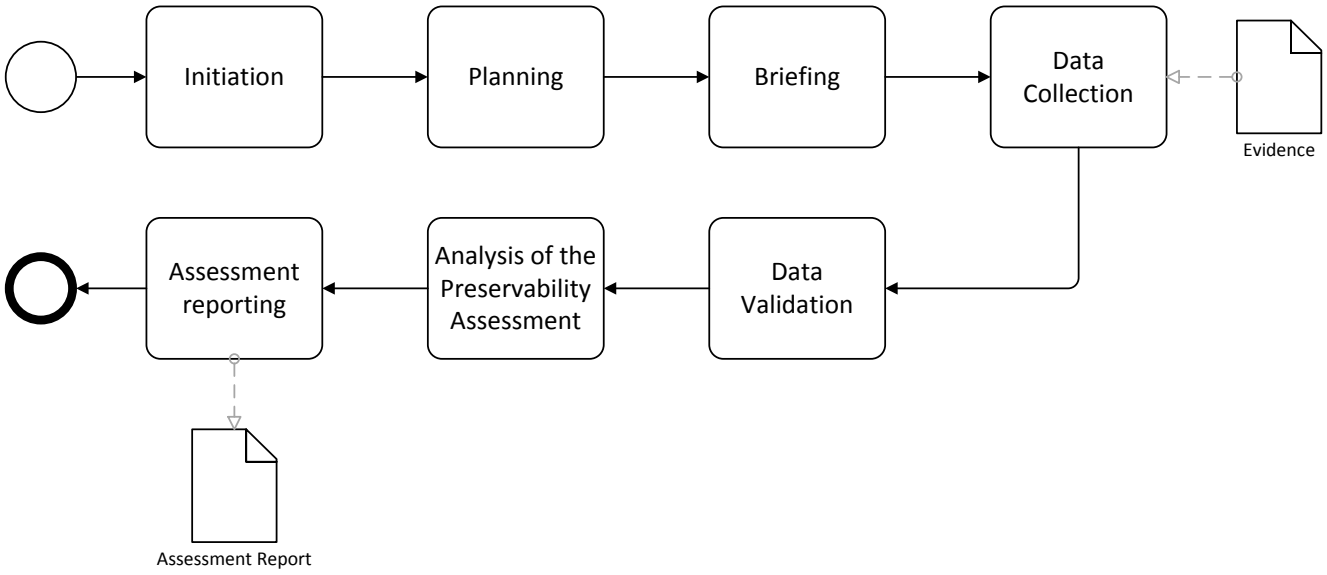


Figure 1: Preservability Assessment Process

3.1 Initiating the Assessment

3.1.1 Overview

The assessment process begins by:

- identifying the stakeholders and defining the purpose of the assessment (why it is being carried out),
- defining what constraints, if any, apply to the assessment,
- identifying any additional information that needs to be gathered,
- choosing the assessment participants and the assessment team and defining the roles of team members,
- defining all assessment inputs and having them approved by the stakeholders.

3.1.2 Tasks

Identify the stakeholders of the assessment.

Select the Assessment Team Leader, who will lead the assessment team and ensure that the persons nominated possess the necessary competency and skills.

Define the assessment purpose including alignment with business goals (where appropriate).

Identify the need for and approve confidentiality agreements (where necessary), especially if external consultants are being used.

Select the Local Assessment Coordinator. The Local Assessment Coordinator manages the assessment logistics and interfaces with the Organization.

Submit Pre-Assessment Questionnaires to the Local Assessment Coordinator. The Pre-Assessment Questionnaires help structure the on-site interviews by gathering information about the Organization and projects.

Establish the assessment team and assign team roles. Normally, the team should ideally consist of two assessors (depending on resource and cost). Assessment team members ensure a balanced set of skills necessary to perform the assessment. The assessment team leader should be a competent assessor.

Define the context. Identify factors in the Organization that affect the assessment process. These factors include, at a minimum:

- the size of the Organization,
- the application domain of the products or services of the Organization,
- the size, criticality and complexity of the products or services,
- the quality characteristics of the products,
- the preservability requirements in terms of quality characteristics of the Organization.

Specify constraints on the conduct of the assessment. The assessment constraints may include:

- availability of key resources,
- the maximum amount of time to be used for the assessment,
- specific Organizations to be excluded from the assessment,
- the minimum, maximum or specific sample size or coverage that is desired for the assessment,
- the ownership of the assessment outputs and any restrictions on their use,
- controls on information resulting from a confidentiality agreement.

Define the goals of the assessment and create the assessment checklist. The goals can be identified and modelled through a goal model (such as i* [4]) which can then be used to create the assessment checklist.

An example of the checklist created for a Civil Engineering Institution assessment regarding the Co-existence quality is shown in Table 1.

Table 1: Co-Existence Assessment Checklist for a Civil Engineering Institution

No.	Compatibility	Evidence
C1	**Co-Existence**	
C1.1	**The system has a historic of compatibility errors which can be traced back to components and maintains an (in)compatibilities list.** *An historic of compatibility errors is very effective to determine the cause of an error as a first attempt, it can be useful to trace errors without much effort. Also, a list of compatibilities and incompatibilities can be used to set up the environment for the system. Example: Two versions of .NET framework installed in the same machine, an outdated driver.*	Logs; Compatibility Errors History Document; (In)compatibilities list; Evidence of continuous update of the (in)compatibilities list; Systems Logs; Document containing the history of errors and possible solutions; Existence of Hardware/Software compatibilities list; Evidence that the Hardware/Software compatibilities list is updated and useful.
C1.2	**There is a mechanism to check for dependencies of system's components and dependencies errors are analyzed by a support team.** *A mechanism to check for (external) components used by a system can help in further installations or exceptions handling, also the analysis of dependencies errors is essential to trace the errors and develop fixes. Example: the use of CUDF (ldd) in LINUX Environments, the use of the registry in Windows environments, dynamic library dependency (otool) in MAC OS.*	Evidence of previous dependency analysis; Evidence of periodic dependency analysis; Evidence of log analysis for co-existence errors; Logs.

Select the assessment participants from within the Organization. The participants should adequately represent the quality characteristics in the assessment scope. As guidance we provide a

set of example organizational roles that can be found across organizations with different backgrounds and different sizes which is based on COBIT 5 [3]. Moreover, these organizational roles are mapped to the characteristics to be assessed in order to get the right people to the assessment. These are provided as guidance, not all organizations have these roles defined in their structure however the role descriptions can help an assessor to find the right people within the organization. The roles and their description are depicted in [5] and the mapping is presented in Table 2. In Table 2 the roles that were not used by any of the quality characteristics were omitted.

Table 2: Mapping of the organizational roles and the preservability assessment characteristics

Id	Quality	Chief Information Officer	Head Architect	Head Development	Head IT Operations	Head IT Administration	Service Manager	Information Security Manager	Privacy Officer
C	**Compatibility**		**x**	**x**	**x**	**x**	**x**		
C1	Co-existence		x	x	x	x	x		
C2	Interoperability		x	x	x	x	x		
P	**Portability**		**x**	**x**	**x**	**x**	**x**		
P1	Adaptability		x	x			x		
P2	Installability		x	x	x		x		
P3	Replaceability		x	x	x	x	x		
M	**Maintainability**		**x**	**x**			**x**		
M1	Modularity		x	x			x		
M2	Reusability		x	x			x		
M3	Analyzability		x	x			x		
M4	Modifiability		x	x			x		
M5	Testability		x	x			x		
S	**Security**	**x**	**x**	**x**		**x**	**x**	**x**	**x**
S1	Confidentiality	x	x	x			x	x	x
S2	Integrity		x	x			x	x	x
S3	Non-repudiation	x	x			x		x	x
S4	Accountability	x	x			x		x	x
S5	Authenticity		x					x	x

Define responsibilities. Define the responsibilities of all individuals participating in the assessment including the stakeholders, assessors, local assessment coordinator and participants.

Identify ownership of the assessment record and the person responsible for approving the assessor logs.

Identify any additional information that the stakeholders requests to be gathered during the assessment.

Review all inputs.

Obtain stakeholders approval of inputs.

3.2 Planning the Assessment

3.2.1 Overview

An assessment plan describing all activities performed in conducting the assessment is developed and documented together with an assessment schedule. Using the project scope, resources necessary to perform the assessment are identified and secured. The method of collating, reviewing, validating and documenting all of the information required for the assessment is determined. Finally, co-ordination with participants in the Organization is planned.

3.2.2 Tasks

Determine the assessment activities. The assessment activities will include all activities described in this documented assessment process but may be tailored as necessary.

Determine the necessary resources and schedule for the assessment. From the scope, identify the time and resources needed to perform the assessment. Resources may include the use of equipment such as overhead projectors, etc.

Define how the assessment data will be collected, recorded, stored, analyzed and presented with reference to the assessment checklist.

Define the planned outputs of the assessment. Assessment outputs desired by the stakeholders in addition to those required as part of the assessment record are identified and described. The output should have in consideration the stakeholder's background, board members or high-level management might want a simple output which shows the present state and which preservability characteristics need improvement. Technical stakeholders might want a detailed feedback on each of the characteristics.

Manage risks. Potential risk factors and mitigation strategies are documented, prioritized and tracked through assessment planning. All identified risks will be monitored throughout the assessment. Potential risks may include changes to the assessment team, organizational changes, changes to the assessment purpose/scope, lack of resources for assessment, confidentiality, priority of the data, and availability of key work products such as documents.

Co-ordinate assessment logistics with the Local Assessment Coordinator. Ensure the compatibility and the availability of technical equipment and confirm that identified workspace and scheduling requirements will be met.

Review and obtain acceptance of the plan. The stakeholders identify who will approve the assessment plan. The plan, including the assessment schedule and logistics for site visits is reviewed and approved.

Confirm the stakeholders' commitment to proceed with the assessment.

3.3 Briefing

3.3.1 Overview

Before the data collection takes place, the Assessment Team Leader ensures that the assessment team understands the assessment input, process and output. The Organization is also briefed on the performance of the assessment.

3.3.2 Tasks

Brief the assessment team. Ensure that the team understands the approach defined in the documented process, the assessment inputs and outputs, and is proficient in using the assessment tool.

Brief the Organization. Explain the assessment purpose, constraints, and process. Stress the confidentiality policy and the benefit of assessment outputs. Present the assessment schedule. Ensure that the staff understands what is being undertaken and their role in the process. Answer any questions or concerns that they may have. Potential participants and anyone who will see the presentation of the final results should be present at the briefing session.

3.4 Data Collection

3.4.1 Overview

Data required performing the assessment is collected in a systematic manner. The strategy and techniques for the selection, collection, analysis of data and justification of the results are explicitly identified and demonstrable. The objective evidence gathered for each criterion assessed must be sufficient to meet the assessment purpose. Objective evidence that supports the assessors' judgment of the criteria compliance is recorded and maintained in the Assessment Record. This Record provides evidence to substantiate the results and to verify compliance with the requirements.

3.4.2 Tasks

Collect evidence of compliance for each criterion.

Record and maintain the references to the evidence that supports the assessors' judgment of the characteristic assessment.

Verify the completeness of the data. Ensure that for each characteristic assessed, sufficient evidence exists to meet the assessment purpose.

3.5 Data Validation

3.5.1 Overview

Actions are taken to ensure that the data is accurate and sufficiently covers the assessment purpose, including seeking information from first hand, independent sources; using past assessment results; and holding feedback sessions to validate the information collected. Some data validation may occur as the data is being collected.

3.5.2 Tasks

Assemble and consolidate the data. For each characteristic, relate the evidence to the criterion.

Validate the data. Ensure that the data collected is correct and objective and that the validated data provides complete coverage of the assessment purpose.

3.6 Analysis of the Preservability Assessment

3.6.1 Overview

For each characteristic, a percentage of compliance is calculated based on the evidence provided by the Organization. Traceability shall be maintained between the objective evidence collected and the percentages calculation.

3.6.2 Tasks

Establish and document the decision-making process used to reach agreement on the results (e.g. consensus of the assessment team or majority vote).

Record the set of percentages for all of the preservability characteristics and calculate the preservability status of the system.

3.7 Reporting the Results

3.7.1 Overview

During this phase, the results of the assessment are analysed and presented in a report. The report also covers any key issues raised during the assessment such as observed areas of strength and weakness and findings of high risk.

3.7.2 Tasks

Prepare the assessment report. Summarise the findings of the assessment, highlighting the key results, observed strengths and weaknesses, and potential improvement actions (if within the purpose of the assessment).

Present the assessment results to the participants. Focus the presentation on defining the state of the preservability characteristics.

Present the assessment results to the stakeholders. The assessment results will also be shared with any parties (e.g. Organization management, practitioners, etc.) specified by the stakeholders.

Finalize the assessment report and distribute to the relevant parties.

Verify and document that the assessment was performed according to requirements.

Assemble the Assessment Record. Provide the Assessment Record to the stakeholders for retention and storage.

Prepare and approve assessor records. For each assessor, records to prove the participation in the assessment are produced. The stakeholders or the stakeholders' delegated authority approves the records.

Provide feedback from the assessment as a means to improve the assessment process.

4. ASSESSMENT RESULTS

The Civil Engineering Institution owns and maintains a system for supporting the process of acquiring and managing information captured from sensors installed in dams, with the objective of

studying the structure behavior and thus prevents any accidents that might happen. Besides managing sensor information, the system, which is called *DamMangement*, is also used for managing the visual inspections, physical models, mathematical models, and technical documents. It also provides data analysis tools such as tabular and chart reports and graphical representation of geo-referenced information.

The *DamMangement* System has the following features:

- Instrumentation: It integrates new observation instruments, supports the dynamic management of new types of instruments, and manages metadata about instruments.
- Transformation process: It manages the instrument specific algorithms to convert raw data into physical actions (results), using instrument metadata properties, such as calibration constants.
- Management of types of observations: It manages geodetic data information, information concerning visual inspections, and data provided by the automatic monitoring systems.
- Data visualization and exploitation: It accesses data through a set of reports designed to support the required types of data analysis, and spatially depicts data using a set of graphics and diagrams.
- Synchronization: It allows the deployment of the system in one or more locations (for example, Civil Engineering Institution and a dam owner) and the corresponding synchronization of data.

This section depicts the detailed assessment results and analysis for the Civil Engineering Institution's *DamMangement* system.

The detailed assessment results are depicted by Figure 2 and Figure 3. These show the results from different levels of detail. The first shows an overview of the different characteristics of preservability, while the second figure shows the in-depth results of each of the sub-characteristics of preservability. Such detailed results are useful for technical stakeholders as it gives a detailed insight on the current state of the different characteristics and sub-characteristics of preservability. The labels for Figure 2 and Figure 3 are shown in Table 3.

Table 3: Charts Label

ID	Name	ID	Name
C	Compatibility	M3	Analyzability
C1	Co-Existence	M4	Modifiability
C2	Interoperability	M5	Testability
P	Portability	S	Security
P1	Adaptability	S1	Confidentiality
P2	Installability	S2	Integrity
P3	Replaceability	S3	Non-repudiation
M	Maintainability	S4	Accountability
M1	Modularity	S5	Authenticity
M2	Reusability		

Figure 2: Compliance Overview Results

In Figure 2, the compliance overview of each of the characteristics is depicted, for the assessed case. The Civil Engineering Institution already has a high degree of security and also a high degree of maintainability, which means that the *DamMangement* system is highly maintainable and secure. Regarding compatibility and portability, the results are lower which might mean that system might not be prepared to be ported into a future environment and that it has not been tested with different components to check for compatibility issues. In Figure 3, we can depict the detailed results for each of the sub-characteristics of preservability. The sub characteristics are now described in increasing detail.

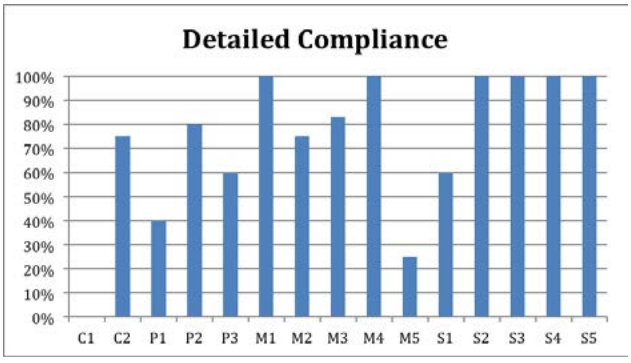


Figure 3: Detailed Compliance Results

4.1 Compatibility Sub-characteristics

The C1 sub-characteristic (Co-existence) could not be found in the *DamMangement* system which means that there is no list of known compatibility and incompatibility issues and there is no mechanism to check for dependencies and errors are not analysed for dependencies issues. The lack of this information might difficult the preservation actions to be taken on the system, since when preserving the system all its dependencies have to be accounted for. The compatibility information is especially important when there is a need for replacing certain components of the system when preserving or redeploying, in the case some original component is/becomes unavailable.

Regarding sub-characteristic C2 (Interoperability), the result attained was almost 80% which shows that *DamMangement* has data transformation mechanisms, it saves all input and output data used in these transformation mechanisms, has documentation about protocols and the interfaces are specified. For preservability, this is a particularly important fact, since data can be migrated to preservation friendly formats, or in future

redeployment scenarios, the system can more easily interoperate with future systems. However, *DamMangement* also uses external or proprietary protocols that might endanger the preservability status of *DamMangement* due to the possibility that these protocols become obsolete.

4.2 Portability Sub-Characteristics

The P1 sub-characteristic (Adaptability) reached 40% of compliance, which means that system already has some degree of adaptability. There is a list of issues concerning the system and software/hardware environments, and the system makes use of open source components, but on the other hands also makes use of proprietary components, which might be troublesome in redeployment scenarios where adaptations to the code have to be made in order to be able to run the system.

The P2 sub-characteristic (Installability) reached 80% of compliance, which shows that *DamMangement* doesn't have any external dependencies in the installation process. The existence of automatic installation packages, the existence of documentation on the resources needed to install the system, and the existence of installation documentation for the system, might contribute to make future installations of the system easier.

The P3 sub-characteristic (Replaceability) reached 60% of compliance, which depicts that in *DamMangement* there is an effort to maintain the system's interface clear so that a replacement would not jeopardise the system functionality. There is also an effort to use the same communication protocols throughout the whole system, when possible. Finally, the system's components are encapsulated in way that facilitates replacement efforts.

4.3 Maintainability Sub-characteristics

In the Maintainability characteristic, the M1 sub-characteristic (Modularity) reached 100% of compliance, which shows that the system has a modular design and that the coupling between modules is low. This is particularly important in preservation and redeployment scenarios where an original component is not available.

The M2 sub-characteristic (Reusability) reached almost 80% of compliance that shows that external interfaces are clearly specified, communication is standardized and the legal regulations in use permit reusability. This contributes for making the redeployment and reuse of a system easier and more trouble free.

The M3 sub-characteristic (Analysability) reached more than 80% of compliance which shows that the system's components have mechanisms which supports analysis, the system is also free from obfuscation techniques, it is implemented according to best practices and standards, is also implemented using popular technology and legal regulations permit analysis.

The M4 sub-characteristic (Modifiability) reached 100%, which shows that *DamMangement* is configurable and that legal regulation allow for modifications to the system, which is crucial so that the system can be configured and modified to adapt to whatever circumstances found in future environments.

The M5 sub-characteristic (Testability) only reached 25%, which shows that the system only allows to be tested without affecting the state of the system. This fact particularly jeopardizes the preservation of the sensor acquisition processes being supported by the system since it might be desirable to test the transformations made to sensor readings in the processes and if the redeployed system is able to provide the same results.

4.4 Security Sub-characteristics

In the Security domain, the S1 sub-characteristic (Confidentiality) reached 60% of compliance that shows that the system allows the specification of access rights to resources, implemented through authorization mechanisms, such as access control lists. It also shows that the system includes encryption mechanisms that might endanger preservability, and manages encrypted information, which can also endanger preservability if the encryption keys are not available in the future.

Regarding the S2 sub-characteristic (Integrity) *DamMangement* reached 100% of compliance that shows that the system includes integrity mechanisms and performs regularly scheduled integrity verifications. This fact is particularly important since it guarantees that the data that is going to be preserved along with the system is not compromised.

The S3 sub-characteristic (Non-repudiation) reached 100% of compliance that depicts that the system has mechanisms for producing records of actions and actively produces records of actions or events on data or components. This fact is particularly important since it ensures that any changes to the system or data are registered, ensuring provenance.

The S4 sub-characteristic (Accountability) reached also 100% of compliance which shows that the system produces records of actions or events on data or components associated with the entities the performed them. This fact is particularly important since it ensures that any changes to the system or data are registered, ensuring provenance.

Finally, the S5 sub-characteristic (Authenticity) reached 100% of compliance that shows that the system has mechanisms for enforcing the authenticity of the entities accessing the system and the system actively enforces the authenticity of the entities accessing the system, thus guaranteeing authentic system components and data.

These results can be used by technical stakeholders to enhance the *DamMangement* system and guarantee that the system is preservable in the future. According to the results, the stakeholders might want to focus the enhancement efforts in the compatibility and portability characteristics.

5. CONCLUSION AND FUTURE OUTLOOK

This paper aimed at the development of a preservability assessment method based on the hypothesis that preservability is a set of systems capabilities originating from a combination of system/software qualities.

An assessment method was proposed to take into account the specifics of each scenario, taking into account the state of the art in assessment checklists and standards on assessment processes. For validating this method, an assessment was performed on the Civil Engineering Institution use case, involving the gathering of the requirements of the stakeholders of the case and the creation of a checklist for assessing preservability in this specific case. The main findings are that the preservability degree of the system described in the industrial case is satisfactory. However, these results can be improved if the documentation concerning some aspects is created and, when existing, that it is kept up to date. Another aspect that can cause preservability to be improved is to keep a registry of compatibility information and performing regular analysis of the compatibility of the system and its components.

The work presented in this paper is not definitive. In fact, it is a proof-of-concept that needs to mature with the application and validation using different scenarios. The example application of the method to the Civil Engineering Institution use case used a checklist where each criterion has a binary evaluation (yes/no), which allows only making limited conclusions. In fact, the desired scenario would be the evaluation of each criterion in a quantitative/qualitative fashion and the creation of a maturity model for preservability against which the evaluation results would be matched. Such scenario is only possible after the application and validation of the method and technique used to several different scenarios which could be used as a benchmark for the creation of the maturity levels.

6. ACKNOWLEDGMENTS

This work was supported by national funds through FCT – Fundação para a Ciência e Tecnologia in context of the pluriannual project PEst-OE/EEI/LA0021/2011, by the project TIMBUS, co-funded by the EU under FP7 under grant agreement no. 269940, and by COMET K1, FFG – Austrian Research Promotion Agency.

7. REFERENCES

[1] ISO/IEC 25010 - Systems and software engineering - Systems and software Quality Requirements and Evaluation (SQuaRE) - System and software quality models, International Organization for Standardization and International Electrotechnical Commission Std. 2010

[2] ISO/IEC 15504 - Information technology — Process assessment, International Organization for Standardization and International Electrotechnical Commission Std. 2004.

[3] IT Governance Institute. COBIT 5 – A business Framework for the Governance and Management of Enterprise IT. 2012.

[4] RWTH i* Guide, [online] http://istar.rwth-aachen.de/tiki-index.php?page=i%2A+Guide&structure=i%2A+Guide (Accessed on 25.02.2013).

[5] IT Governance Institute. *COBIT 5 – Enabling Processes.* 2012.

An Analysis of Contemporary JPEG2000 Codecs for Image Format Migration

William Palmer
British Library
96 Euston Road
London, United Kingdom
william.palmer@bl.uk

Peter May
British Library
96 Euston Road
London, United Kingdom
peter.may@bl.uk

Peter Cliff
British Library
96 Euston Road
London, United Kingdom
peter.cliff@bl.uk

ABSTRACT

This paper presents results of an analysis of different implementations of the JPEG2000 standard, specifically part 1: JP2, an image format that is currently popular within the digital preservation community. In particular we are interested in the effect different JPEG2000 codecs (encoders and decoders) have on image quality in response to lossy compression. We focus on three main codec libraries for analysis - Kakadu, JasPer and OpenJPEG - migrating 932 TIFF newspaper images to lossy JPEG2000 files using 2:1 and 4:1 compression ratios, and monitor image quality using PSNR. We look at the combination of encoder/decoder pairs and find that using OpenJPEG for both gives the best quality results, albeit with the slowest execution time. We also find that in some circumstances, particularly when a JasPer encoder is used, in order to retain image quality of the decoded image, the best choice of decoder may not be the same codec used to create the JPEG2000; based on these results, the encoding library is therefore recommended technical preservation metadata to retain.

Keywords

JPEG2000, TIFF, migration, codec, PSNR, image quality, generational loss

1. INTRODUCTION

The British Library, as a memory institution, holds large quantities of digital content, including over 2 million files produced from the JISC funded British Newspaper digitisation projects [1]. The number of files within our digital collections is ever increasing and these need to be cost-effectively preserved to ensure long-term access to these images.

Storing this collection as TIFF files would require a significant amount of storage just to preserve the masters alone. JPEG2000 presents an alternative file format for digital images, that has a number of advantages for preservation, such as reduced storage costs compared with the traditional TIFF master files and the ability to contain both master and lower-resolution access copies within a single file (compared with TIFF masters which typically require a separate access copy - PNG or JPEG - to also be kept) [2].

JPEG2000 files can be compressed in either of two ways; losslessly, where a bit-identical copy of the data is maintained and can be retrieved, and lossy, where an exact copy of the data is not maintained and some fidelity is lost. In general, lossy compression results in smaller image files, but will suffer from information loss resulting in image artefacts such as blurring, as is especially the case with high compression ratios.

Putting aside the merits of preserving image files using lossy compression, what is unclear is whether all JPEG2000 codecs perform to the same quality. Do all codecs produce the same quality of file given the same settings, or are there consistent variations between them? Ebrahimi et. al. [3] considered the effects of compression ratios on perceived image quality using three JPEG2000 codecs (JasPer, Kakadu and IrfanView). Their results suggested JasPer performed the better of the three (it is unclear if this result is statistically significant), however this operated on a small sample (29) of small files (768x512 pixels) from the LIVE Image Quality database [4], using a large range of compression ratios (from 2:1 - 300:1), and focussed primarily on the metric for determining image quality rather than tool performance itself (quality of result, speed, etc.). In contrast, a sample of the JPEG2000 files from our newspaper collection were images of sizes 4672x5944 pixels and 8320x9568 pixels, with applied compression ratios at the low end of the spectrum, around 2:1 (although, as supported by JPEG2000, a number of compression levels have been defined to provide numerous quality-level images within the one file, for example, for access).

If we consider the preservation process for migrating an image to JPEG2000 and then accessing it, even for quality assurance purposes, then two distinct uses of the codec are required. The original image must first be encoded using a JPEG2000 codec to create a JP2 file, then accessed through a decoding step with a JPEG2000 codec. However the codec used for these two steps does not have to be the same.

From a preservation perspective, it would therefore be useful to understand the effects of various codecs, combinations of encoding/decoding codecs, and codec settings on the resulting files, thus providing evidence to enable appropriate tool selection within a migration workflow. This paper presents our initial analysis results of such codec use for a

single migration, before looking at an extension considering the effects of lossy migration on subsequent migrations.

To put the extension into context, we expect to provide access to our files for many hundreds of years into the future. Although Gollins [5] promotes a parsimonious "rule-of-thumb" approach to preservation, "using only the minimum necessary intervention to secure our digital heritage for the next generation", no guarantees can be made about the absence of future migrations. Of concern when migrating files between lossy formats is digital generational loss - the increasing loss of information with each migration. Although it is fairly logical to conclude that lossy compression, having resulted in information loss, will cause ever increasing degradation in subsequent lossy migrations, to what extent will this degradation be? How quickly will it degrade? And is the amount of degradation affected by codec choice?

This paper starts by looking at the JPEG2000 codec libraries available, mentioning details of profiles, specifically with respect to compression rates, and indicating choices made for the experimental work carried out. Section 3 details the methodology of our single-migration codec analysis, as well as the generational loss extension. Results are presented in Section 4, with general conclusions and ideas for future work presented in Section 5.

2. USE OF JPEG2000

Whilst the British Library and other memory institutions [2] now utilise JPEG2000 files for preservation, their utilisation outside of these organisations would appear low. As an indication of the mainstream use of JP2 files, a recent search for "image/jp2" content type over the UK Web Domain Dataset Format Profile (1996-2010)[6] found only 53 files with some identification as "image/jp2". To put this further into context, a similar search for "image/jpeg" returned over 153 million files.

This apparently low uptake in the wider world may, in itself, present a preservation risk through lack of "high-quality" tool support [7]. Irrespective of whether this is an issue or not, there are several commercial and open-source libraries currently available; the question this paper starts to address is how do these compare?

2.1 JPEG2000 Libraries

There are several JPEG2000 codec libraries currently available, including:

- Kakadu[1], last updated January 2013. Commercial.
- OpenJPEG[2], last updated November 2012. BSD 3-Clause license.
- JasPer[3][8], last updated January 2007. License based on MIT license.
- JJ2000[4], last modified date is November 2009 (note: the actual last modification is possibly much before that). License unclear.
- FFMPEG[5], last updated December 2012. (L)GPL license.
- Other commercial codecs: Aware, LuraTech, LeadTools & J2K Codec

For this analysis we chose to focus on the first three tools: Kakadu because it appears to be the codec of choice for large institutions; OpenJPEG as it is an open source codec that is actively maintained; and JasPer as it is the JPEG2000 codec widely used by other open source projects.

Conversely, we decided not to use JJ2000 due to its lack of recent development activity and mainstream use; FFMPEG for similar reasons; nor other commercial tools as we had no readily-available access. These codecs are in consideration for future research (see Section 5.1).

2.2 Profiles

A profile specifies desired image properties, such as compression (reversible or irreversible) and quality layers (with associated compression rates). From this, appropriate control settings to be used by a codec to create a JPEG2000 image can be derived. Significant investment seems to have been placed in trying to identify an appropriate level of compression whilst maintaining an acceptable quality of archival and production masters. Techniques mentioned in [2] formed around using either objective judgements from human observers to determine a visually lossless compression ratio, or by taking a minimal loss approach through compression with a maximum bit rate[6], i.e. all data is retained but there is minimal loss through rounding errors introduced by floating point transforms and from quantization. These approaches have resulted in similar findings regarding appropriate levels of compression, however it should be noted that the necessary compression settings needed to achieve visually lossless images is dependent on the images themselves [2].

The National Digital Newspaper Program (NDPD), for example, decided on an 8:1 compression ratio for JPEG 2000 production masters[7] as a compromise between file size and image quality, although 4:1 and 6:1 were judged to be visually lossless [9].

The Wellcome Trust's digital library use an iterative approach to determining compression rates across a collection (increasing compression until artefacts are observed, then stepping back), but have found a 10:1 compression ratio works well for books and 8:1 for archive collection material [2].

The British Library's recommended JPEG2000 profile for use in mass digitisation projects, in particular for our newspaper digitisation, is detailed in [10]. This specifies 12 quality layers, with compression levels starting at a minimally lossy rate. Whilst Kakadu and OpenJPEG can encode images according to this profile, JasPer cannot as it cannot use different precinct sizes. Additionally, there seems to be a bug in OpenJPEG v2.0.0 when coder bypass is used[8].

To make a comparative analysis of the codecs we chose a profile that could be encoded by all three chosen coders based on the British Library's recommended profile [10]. This is shown in Table 1.

2.3 Automated Codec Comparison

Much research has been done on Image Quality Algorithms (IQA) for measuring visible image quality [3, 11], however as Buckley [2] notes, currently "there is simply no

[1] www.kakadusoftware.com
[2] www.openjpeg.org
[3] www.ece.uvic.ca/~frodo/jasper/
[4] code.google.com/p/jj2000/
[5] git.videolan.org/?p=ffmpeg.git;a=blob;f=libavcodec/j2kenc.c

[6] Compressed bit rate is the ratio of compressed image data size to the image width and height[2]
[7] NDPD use uncompressed TIFF files for preservation masters.
[8] code.google.com/p/openjpeg/issues/detail?id=209

Table 1: Test JPEG2000 profile

File format	JP2
Transformation	9-7 irreversible (lossy)
Progression order	RPCL
Tiling	none
Levels	6
Precinct size	all 256x256
Quality layers	one
Code block size	64x64
Coder bypass	no

substitute for a human observer". Despite this, for the large collections held by us, human observation over the entire collection is simply not practical; more automated means are required.

One such IQA is the peak signal-to-noise ratio (PSNR) which gives a numerical value of the errors introduced by a lossy image encode, on a logarithmic scale, measured in decibels. A higher value indicates a better ratio of signal-to-noise, and provides some indication as to the quality of the image.

As a metric, PSNR is considered not to match well to perceived visual quality [12]. As part of our investigation we also tested use of a tool that calculated SSIM (Structural Similarity) as a metric for image file format migration quality. We found that SSIM was less sensitive to changes in the image than PSNR and was good for correlating *similar* images, such as resized images or those with added noise. However, the runtime of that tool was longer than ImageMagick's PSNR comparison. From our analysis the better metric of the two for image *identicalness* (as would be expected from a migration) was PSNR as it was faster and more sensitive to small changes in an image, thus giving greater assurance of success. We therefore opted to use PSNR as metric of image quality for this work.

3. METHODOLOGY

Migrating a TIFF to JPEG2000 and then viewing (or performing image analysis on) the resulting file requires both an encode and decode step. An assumption is typically made that the same codec should be used for both, but that does not need to be the case. The base experiment compares the 9 different combinations of encoder-decoder pairings possible with three libraries (Kakadu 7.1, OpenJPEG 2.0 and JasPer 1.900.1-13), see Table 4. This work is then extended to consider the effects of generational loss, i.e. how multiple migrations (encode-decode cycles) affect image quality.

3.1 Dataset

The input dataset used was 932 greyscale TIFF original masters from the British Library's JISC1 newspaper collection, totalling 26GB. Images from this sample averaged 51.0 megapixels, with a minimum of 21.1 megapixels and maximum of 93.4 megapixels.

3.2 Data Preparation

JasPer cannot take TIFF as an input format, therefore to make the migration experiments fairer, the TIFF input files are first converted to PNM (Portable Any Map) files using ImageMagick's *tifftopnm*, version 6.6.9.7-5ubuntu3.2.

3.3 Comparison Approach

A program was written to generate shell scripts that performed the following steps on a PNM file, for each encoder-decoder pairing. For each encoder, the appropriate command line listed in Section 3.4 was used to obtain JPEG2000 images meeting the desired profile. The steps are:

1. **Migrate:** Use the specified encoder and decoder to convert the PNM to a JP2 and then back to a PNM file (this is repeated as necessary for generational loss test);
2. **Validate Profile:** Extract information from the JP2 file using Jpylyzer, for later validation against the desired profile (specified in Section 2.2);
3. **Calculate PSNR:** Use ImageMagick to calculate the PSNR of the original TIFF file versus the migrated output PNM file (this comparison with the original is repeated after each migration for generational loss analysis), storing results in a CSV file;
4. **Consolidate Results:** Create a zip file that collects outputs from the above

This program was wrapped in a Hadoop MapReduce program so that, for each input TIFF file, firstly the tool is executed to generate the necessary shell scripts, and then these generated scripts are executed. A separate program was produced that extracted information from all the run outputs and produced aggregated CSV files.

Using a compression ratio of 2:1, runs were made consisting of ten generations of encode-decodes for each encoder-decoder pair, with, as per the methodology above, PSNR calculated between the original file and the output PNM after each generation. For these runs, it was found JasPer did not use all the space available to it - its compressed files were consistently more compressed than the requested compression ratio. Consequently, to enable all three encoders to produce output files of the same size and compression ratio, a further run was made using a 4:1 compression ratio. Nearly all files for each coder were within a few bytes of each other - see Table 3 - however, for some files JasPer still over-compressed them.

3.4 Encoder Command Line Parameters

The command line parameters which generate images conforming to the profile specified in Section 2.2, for each codec library, are shown in Table 2.

Table 2: JPEG2000 command lines for each library

Kakadu	Creversible=no Corder=RPCL Clevels=6 Cprecincts={256,256} Cblk={64,64}
OpenJPEG	-I -p RPCL -n 7 -c [256,256],[256,256],[256,256],[256,256], [256,256],[256,256],[256,256] -b 64,64
JasPer	-T jp2 -O mode=real -O prg=rpcl -O numrlvls=7 -O prcwidth=256 -O prcheight=256 -O cblkwidth=64 -O cblkheight=64

Note that the number of levels requested differs between Kakadu and the other tools, however, analysis of the outputs using Jpylyzer[13] shows these results to be equivalent. This discrepancy could be due to codec authors' different interpretations of the specification.

4. RESULTS

4.1 Exact Re-generation of Compressed Files

On each test run the JPEG2000 files were recreated (encoded) three times by each encoder. Each set of encodes produced identical files according to their MD5 checksum, suggesting that there is no variation produced by each library during encoding.

4.2 Compression Ratio

Kakadu and OpenJPEG's encoders routinely meet the compression ratio asked of them. There was a difference with the JasPer encoder, in that if it did not need all the space afforded by the specified compression ratio, it encoded at a higher compression ratio, producing a smaller file. This was also expected of the other codecs at lower compression ratios but was not apparent, from our results, at 2:1 compression.

4.3 File size

At 2:1 compression, the mean file size difference between the Kakadu and OpenJPEG encoders was 0.04% ±0.03% (7467 bytes ±5938 bytes). As already noted, JasPer encoded smaller files than the requested compression ratio would entail.

At 4:1 JasPer compressed files that were closer to the other encoders, see Table 3. However, as can be seen in Figure 2, at a requested 4:1 compression, JasPer was not always able to fully utilise the compression ratio.

Table 3: Mean difference between compressed file sizes at 4:1 compression for encoder-decoder pairs

Kakadu - OpenJPEG	OpenJPEG - JasPer	Kakadu - JasPer
186 ±106 bytes (0.002% ±0.001%)	747093 ±1102190 bytes (5.4% ±8.3%)	747277 ±1102201 bytes (5.4% ±8.3%)

4.4 Single Migration Image Quality

The results from the first encode-decode cycle for each encoder-decoder pairing are shown in Figure 1 and Figure 2, with the corresponding mean average PSNR (and standard deviation) shown in Table 4.

Table 4: Mean average PSNR for each encoder-decoder pair at 2:1 and 4:1 Compression rates

Encoder-Decoder	Mean Average PSNR (2d.p)	
	2:1 rate	4:1 rate
jasper-jasper	47.81 ±0.50dB	46.96 ±1.54dB
jasper-opj20	49.56 ±0.32dB	47.78 ±2.12dB
jasper-kakadu	49.45 ±0.32dB	47.69 ±2.08dB
kakadu-jasper	50.34 ±0.35dB	47.20 ±2.03dB
kakadu-opj20	54.17 ±0.41dB	48.12 ±2.44dB
kakadu-kakadu	54.22 ±0.43dB	48.13 ±2.45dB
opj20-jasper	52.62 ±0.37dB	48.23 ±2.45dB
opj20-opj20	**55.02 ±0.52dB**	**48.30 ±2.51dB**
opj20-kakadu	54.58 ±0.48dB	48.23 ±2.48dB

They show the OpenJPEG encoder with OpenJPEG decoder to produce the highest average PSNR for both 2:1 and 4:1 compression rates, at 55.02dB (2d.p.) and 48.30dB (2d.p.) respectively. The next highest, with a slight drop in PSNR, is the OpenJPEG-Kakadu pairing (54.58dB and 48.23dB for 2:1 and 4:1 respectively), followed by the Kakadu encoder with either using OpenJPEG or Kakadu for decoding (both showing 54.2dB and 48.1dB to 1d.p. for 2:1 and 4:1 respectively).

Our results indicate that files encoded with JasPer and decoded using any of the tools tend to result in a lower PSNR ($< 50dB$ for 2:1 and $< 48dB$ for 4:1) than when using other libraries for encoding. Using OpenJPEG or Kakadu as the decoder for such encoded files gives slightly better average PSNR results than using JasPer as the decoder (approx 49.5dB as opposed to 47.8dB for 2:1 compression; approx 47.7dB as opposed to 47dB for 4:1 compression).

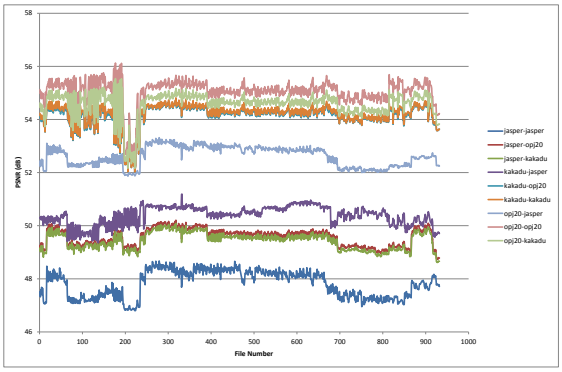


Figure 1: PSNR for first encode-decode for each encoder-decoder pair at 2:1 compression

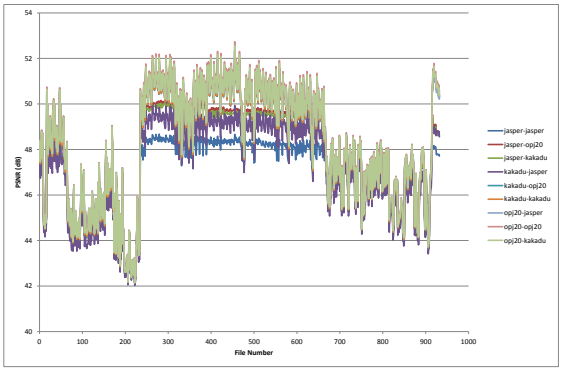


Figure 2: PSNR for first encode-decode for each encoder-decoder pair at 4:1 compression

A statistical analysis of the mean average PSNRs for each encoder-decoder pair (at 2:1 compression) showed that there is a statistically significant difference in the mean average PSNR, at 0.05 level, between all combinations of encoder-decoder pairs, apart from between Kakadu-OpenJPEG and Kakadu-Kakadu (the orange and hidden blue line 3rd and 4th from the top in Figure 1). This is congruent with the results in Table 4, which show the mean average PSNRs for these two pairings to be almost identical.

The difference in the PSNR means between the JasPer-OpenJPEG and JasPer-Kakadu pairs (green and red lines 2nd and 3rd from the bottom in Figure 1) was only just statistically significant. Again, this is reflected in the closeness of mean PSNRs seen in Table 4.

For the 4:1 compression ratio, shown in Figure 2, the statistical analysis of the difference in average PSNR between each encoder-decoder pair showed that there is a statistical

significance, at 0.05 level, between JasPer-JasPer (blue line) and all other encoder-decoder pairs, apart from Kakadu-JasPer (purple line). This corresponds to a 0.24dB difference in mean average PSNR, and so the lack of statistical significance is unsurprising. Similarly, there is a statistical significance in the mean PSNR between Kakadu-JasPer (purple line) and all other encoder-decoder pairs, apart from JasPer-JasPer.

For the remaining differences between pairs, the JasPer-OpenJPEG or JasPer-Kakadu pairings compared against any other combination typically showed low levels of significance (some showed no significance, for example JasPer-OpenJPEG vs Kakadu -Kakadu), reflective of the small differences in mean PSNR showed in Table 4. All other combinations have even lower differences in mean PSNRs which are not statistically significant.

4.5 Generational loss

Ten encode-decode cycles were run for each file with each encoder-decoder pair. This was to test how the encoder-decoder pairs coped with generational loss through repeated migrations of lossy-encoded images.

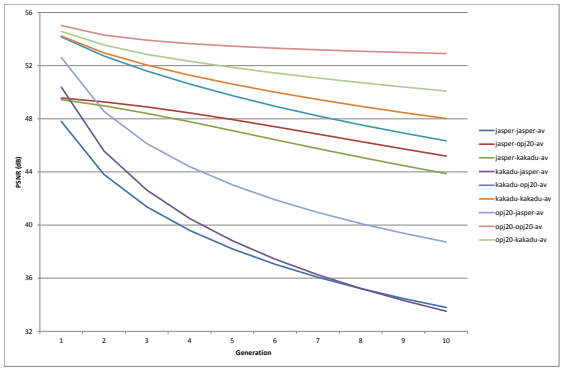


Figure 3: PSNR showing ten generations of 2:1 encode-decode for each encoder-decoder pair

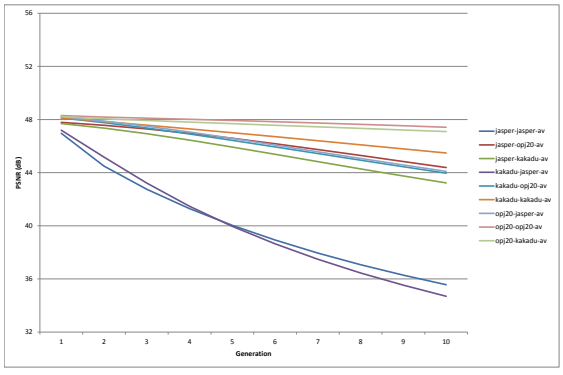


Figure 4: PSNR showing ten generations of 4:1 encode-decode for each encoder-decoder pair

Figures 3 and 4 show the mean average PSNRs after each generation, for 10 encode-decode cycles. The results are the average across all 932 files.

For both 2:1 and 4:1 compression ratios, all libraries show some signs of PSNR degradation, however some codecs suffer less than others. The OpenJPEG-OpenJPEG pairing is consistently the best (for both compression ratios) with only an average drop of 2.11dB (2d.p.) in PSNR for 2:1 compression and 0.88dB (2d.p.) for 4:1, over 10 generations.

At 2:1 compression, using JasPer as a decoder in combination with any other encoder results in the largest drops in PSNR - at least 13.89dB (2d.p.; OpenJPEG-JasPer). In particular though, the Kakadu-JasPer and JasPer-JasPer pairings both show the largest drops in PSNR for both compression rates.

4.6 Pixel differences vs original

As another means to gauge image quality, we used ImageMagick to count the number of pixels changed between the original image and the encode-decoded images. At 2:1 compression the maximum number of pixels that were more than 1% different to the original pixels, for Kakadu and OpenJPEG encoded files using Kakadu and OpenJPEG decoders, was 549. The average total number of pixels with an absolute difference to the original was between 20-24% of image pixels. JasPer figures are not quoted because, as previously described, at 2:1 compression JasPer tends to over-compress.

At 4:1 compression the percentage differences between encoder-decoder pairs are close. The overall mean of the average differences for each codec pair suggests approximately 2% of pixels are greater than 1% different to those in the original image. The same calculation for absolute differences in pixels gives an average of 55% of pixels being different.

4.7 Execution speed

A record was kept of execution speed for the generational loss encode-decode process, measuring the initial conversion from TIFF to PNM, using ImageMagick's *tifftopnm*, followed by ten encode-decode cycles. Files were decoded by the decoder directly to PNM, ready for the next encode.

Table 5 shows the average speed per encode-decode cycle for each encoder-decoder pair. Using Kakadu to encode and decode images is the quickest, using OpenJPEG is the slowest. Interestingly, using OpenJPEG as the encoder and/or decoder tends to result in longer encode-decode cycles.

Table 5: Mean execution speed of encoder-decoder pairs, per encode-decode cycle

Encoder-Decoder	Speed (s) at 2:1	Speed (s) at 4:1
Kakadu-Kakadu	16	12
Kakadu-JasPer	24	23
JasPer-Kakadu	25	24
JasPer-JasPer	32	32
Kakadu-OpenJPEG	44	34
JasPer-OpenJPEG	48	46
OpenJPEG-Kakadu	75	68
OpenJPEG-JasPer	78	78
OpenJPEG-OpenJPEG	103	90

5. CONCLUSION

1. For our test images and codec settings, the best quality encoder-decoder pair was OpenJPEG-OpenJPEG. The next best, where a different encoder was used, was the Kakadu-Kakadu pair, at approximately 1dB lower than OpenJPEG-OpenJPEG at 2:1 compression. It is worth noting that the execution speed of those pairs is both the fastest (Kakadu-Kakadu) and slowest (OpenJPEG-OpenJPEG). The file size difference between these encoders at 2:1 and 4:1 was low.

2. It appears that when a specific compression ratio is requested of the JasPer encoder, it compresses up to that ratio. If the encoder is unable to use all the space afforded to it by the compression ratio, the resulting file will have a higher compression ratio. This is readily apparent in the 2:1 compression test, and apparent in the 4:1 compression test.
3. From the results of the first encode-decode cycle for the codec pairs, we see that the JasPer decoder does not produce as high quality output as other decoders, given the same input file. Our results show the JasPer decoder was up to approximately 4dB worse quality than other decoders (e.g. Kakadu-Kakadu vs Kakadu-JasPer at 2:1 compression).
4. The first encode-decode at a higher compression ratio of 4:1, where the JasPer encoder was able to compress to the requested ratio, produced files much closer to the requested ratio than at 2:1. Results for these cases showed that only the JasPer-JasPer and Kakadu-JasPer pairs performed notably worse than other pairs. The average PSNR values for those two pairs were statistically significant from the other encoder-decoder pairs.
5. The lower quality output of the JasPer decoder became apparent in the generational loss tests, where use of a JasPer decoder, especially in conjunction with a JasPer or Kakadu encoder, produced notably larger drops in PSNR. When Kakadu or OpenJPEG decoders were used, the PSNR drop was less severe; they produced better quality decoded images from the same compressed JP2.
6. The JasPer decoder is unable to decode any of the test JP2 files to the same quality as the other decoders, including JPEG2000 files encoded by its own encoder. If quality of decoded files is important it may be advisable to decode JPEG2000 images with a decoder known to be good for the encoder rather than using an unknown/lower quality built-in decoder. For digital preservation, this indicates an importance in understanding the library used to create the JPEG2000 file.

5.1 Future work

There are many potential avenues to extend this work, but some notable areas we feel warrant further work include:

- Repeating the testing with different collections, for example; photographs
- Using different codecs and settings, for example; using Kakadu's *-precise* flag to increase precision
- Using different encoding profiles; tiles vs precincts, compression ratios, compression layers
- Using different quality metric(s)

6. ACKNOWLEDGEMENTS

This work was partially supported by the SCAPE project (www.scape-project.eu). The SCAPE project is co-funded by the European Union under FP7 ICT-2009.4.1 (Grant Agreement number 270137).

7. REFERENCES

[1] *The JISC Digitisation Programme: Overview of Projects*. URL: http://www.jisc.ac.uk/media/documents/programmes/digitisation/digitisation_v2_overview_final.pdf (Last accessed: 04/26/2013).

[2] Robert Buckley. *Using Lossy JPEG 2000 Compression for Archival Master Files*. Library of Congress, Mar. 12, 2013. URL: http://www.digitizationguidelines.gov/still-image/documents/JP2LossyCompression.pdf (Last accessed: 04/25/2013).

[3] Farzad Ebrahimi, Matthieu Chamik, and Stefan Winkler. "JPEG vs. JPEG 2000: an objective comparison of image encoding quality". In: *Proceedings of SPIE*. Vol. 5558. 2004, pp. 300–308. URL: http://stefan.winklerbros.net/Publications/adip2004.pdf (Last accessed: 04/26/2013).

[4] H. R. Sheikh et al. *LIVE image quality assessment database (2003)*. 2003. URL: http://live.ece.utexas.edu/research/quality/ (Last accessed: 04/26/2013).

[5] Tim Gollins. "Parsimonious preservation: preventing pointless processes!" In: *Online Information*. 2009, pp. 75–78. URL: http://www.nationalarchives.gov.uk/documents/parsimonious-preservation.pdf (Last accessed: 04/26/2013).

[6] *Format Profile JISC UK Web Domain Dataset (1996-2010)*. DOI: 10.5259/ukwa.ds.2/fmt/1.

[7] *Is JPEG-2000 a Preservation Risk?* Jan. 28, 2013. URL: http://blogs.loc.gov/digitalpreservation/2013/01/is-jpeg-2000-a-preservation-risk/ (Last accessed: 04/24/2013).

[8] M.D. Adams and F. Kossentini. "JasPer: a software-based JPEG-2000 codec implementation". In: *Image Processing, 2000. Proceedings. 2000 International Conference on*. Vol. 2. 2000, 53–56 vol.2. DOI: 10.1109/ICIP.2000.899223.

[9] Robert Buckley and Roger Sam. *JPEG 2000 Profile for the National Digital Newspaper Program*. Apr. 27, 2006. URL: http://www.loc.gov/ndnp/guidelines/docs/NDNP_JP2HistNewsProfile.pdf (Last accessed: 04/25/2013).

[10] *The British Library JPEG 2000 Profile for Bulk Digitisation*. URL: http://www.digitizationguidelines.gov/still-image/documents/Martin.pdf (Last accessed: 04/26/2013).

[11] Thien Phan, Phong Vu, and Damon M. Chandler. "On the Use of Image Quality Estimators for Improved JPEG2000 coding". In: Asilomar Conference on Signals, Systems, and Computers (2012). 2012. URL: http://vision.okstate.edu/pubs/asilomar_2012.pdf (Last accessed: 04/26/2013).

[12] Zhou Wang et al. "Image Quality Assessment: From Error Visibility to Structural Similarity". In: *IEEE Transactions on Image Processing* 13.4 (Apr. 2004). URL: https://ece.uwaterloo.ca/~z70wang/publications/ssim.pdf (Last accessed: 04/26/2013).

[13] David Tarrant and Johan Van Der Knijff. "Jpylyzer: Analysing JP2000 files with a community supported tool". 2012. URL: http://eprints.soton.ac.uk/341992/.

Managing and Transforming Digital Forensics Metadata for Digital Collections

Kam Woods, Alexandra Chassanoff, Christopher A. Lee
University of North Carolina - School of Library and Information Science
216 Lenoir Drive, CB #3360, 100 Manning Hall
Chapel Hill, NC 27599-3360
(919) 962-8366
{kamwoods, achass, callee}@email.unc.edu

ABSTRACT
In this paper we present ongoing work conducted as part of the BitCurator project to develop reusable, extensible strategies for transforming and incorporating metadata produced by digital forensics tools into archival metadata schemas. We focus on the metadata produced by open-source tools that support Digital Forensics XML (DFXML), and we describe how portions of this metadata can be used when recording PREMIS events to describe activities relevant to preservation and access. We examine open issues associated with these transformations and suggest scenarios in which capturing forensic metadata can support digital curation goals by establishing clear documentation of integrity and provenance, tracking events associated with pre-ingest and post-ingest forensic processing, and providing specific evidence of authenticity.

Categories and Subject Descriptors
H.3.7 [**Information Storage and Retrieval**]: Digital Libraries – *collection, dissemination, systems issues*

General Terms
Documentation, Reliability, Security, Legal Aspects.

Keywords
Digital forensics, disk imaging, preservation metadata, DFXML, interoperability, BitCurator.

1. INTRODUCTION
The process of preserving data encoded on digital media by extracting a bit-identical disk image has many advantages [14,15,23], but it also presents unique technical and organizational challenges. Many of these challenges are related to the metadata produced during acquisition, processing, and archival management of born-digital materials. The challenges often arise from working with implementations of metadata schemas that are *document-centric*; that is, schemas which have been designed primarily to accommodate the acquisition, analysis, and transformation of individual files (e.g., Microsoft Word documents or TIFF images). A disk image, in contrast, may contain hundreds, thousands, or even millions of files with many potential internal dependencies [23]. The disk image itself may not always be the final preservation target, but capturing and describing information about the internal structure and any potential dependencies is an important aspect of supporting ongoing preservation activities, as well as meaningful access and use.

Forensic analysis of disk images often produces large quantities of metadata. Much of this forensic metadata is initially reported at a very low level; for example, as patterns identified at various offsets into the raw bitstream. These reports may be transformed using a variety of intermediate procedures in order to generate derived metadata for specific tasks: retention within an Archival Information Package; storage within a database in preservation and access systems; or to support archival lifecycle processes.

In the following sections we describe specific metadata elements that can be extracted or derived using the BitCurator environment, and our evolving approach to mapping these items to archival metadata standards. In this paper, the preservation metadata target we focus on is PREMIS.

2. ACQUIRING FORENSIC METADATA IN BITCURATOR
The BitCurator Project is a collaborative effort led by the School of Information and Library Science at the University of North Carolina at Chapel Hill and the Maryland Institute for Technology in the Humanities at the University of Maryland. BitCurator aims to address two fundamental needs and opportunities for collecting institutions: (1) integrating digital forensics tools and methods into the workflows and collection management environments of libraries, archives and museums; and (2) supporting (potentially mediated) public access to forensically acquired data [16].

We are developing and disseminating a suite of open source tools. These tools are currently being developed and tested in a Linux environment. The majority of the software on which they depend can be compiled for Windows environments (and in most cases are currently distributed as both source code and Windows binaries), or runs in a cross-platform interpreter. We intend the majority of the development for BitCurator to support cross-platform use of the software. We are freely disseminating software developed by BitCurator under an open source (GPL, Version 3) license. All other software packaged within the BitCurator environment is distributed in accordance with the terms of the

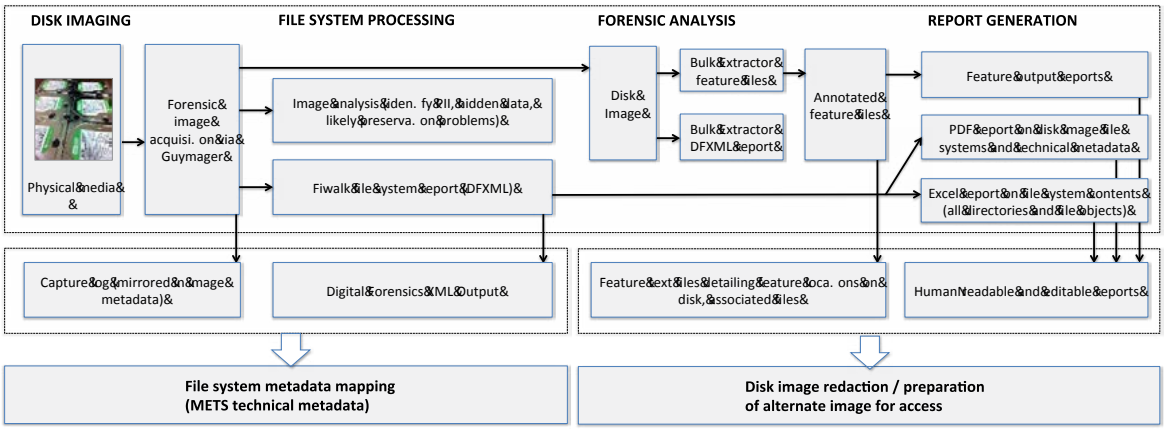


Figure 1: Overview of BitCurator disk image processing, metadata extraction, and redaction.

```
GUYMAGER ACQUISITION INFO FILE
==============================
Version: 0.4.2-2

Compilation timestamp: 2010-02-08-14.45.08
Compiled with: gcc 4.4.3
libewf version: 20100226
libguytools version: 1.1.1
Linux device: /dev/sdc
Device size: 2000398934016 (2.0TB)
Image path and file name: /media/DataVolume2/SampleImage
Info path and file name: /media/DataVolume2/SampleImage
Image format: Advanced forensic image - file extension is .aff

MD5 hash: d3948773eea011ffa559009881da8a8e
MD5 hash verified   : --
SHA256 hash:
5859b189298ee319c291a9286326080aa1b60ab41f632adbaf6d22a
e3c7f3444
```

Figure 2: Sample log metadata produced by Guymager during disk image acquisition.

capabilities of the physical device and would only be used in specialized circumstances. Essential details of the image capture (in the original DFXML) may be wrapped as technical metadata capture), technical details of the processing environments, and MD5 and SHA256 checksums.

2.2 File system metadata

Information about the file system(s) contained within a disk image can be extracted using the *fiwalk* tool integrated into the current version of The Sleuth Kit, which is itself incorporated into the BitCurator environment. Output of *fiwalk* incorporates Dublin Core tags to identify the *creator* of the DFXML file (*fiwalk*, along with technical details on the environment in which it was run) and the *source* (the disk image file that was scanned). Note that these tags should not be treated the same as archival descriptive metadata. They are more accurately incorporated as *technical metadata* corresponding to an intermediate event (analysis of the file system(s) contained within the disk image).

A partial example of the DFXML output produced by *fiwalk* is shown in Figure 3. This section was extracted from the head of the DFXML file. Note the inclusion of technical details corresponding to the capture environment, a start timestamp in

```
<?xml version='1.0' encoding='UTF-8'?>
<dfxml version='1.0'>
  <metadata

xmlns='http://www.forensicswiki.org/wiki/Category:Digital_Fo
rensics_XML'
  xmlns:xsi='http://www.w3.org/2001/XMLSchema-instance'
  xmlns:dc='http://purl.org/dc/elements/1.1/'>
    <dc:type>Disk Image</dc:type>
  </metadata>
  <creator version='1.0'>
    <program>fiwalk</program>
    <version>4.0.2</version>
    <build_environment>
      <compiler>GCC 4.7</compiler>
      <library name="afflib" version="3.7.1"/>
      <library name="libewf" version="20130303"/>
    </build_environment>
    <execution_environment>
      <command_line>fiwalk -f -X
/media/kamwoods/DataVolume2/SampleImage.xml
SampleImage.aff</command_line>
      <start_time>2013-03-29T16:46:13Z</start_time>
    </execution_environment>
  </creator>
  <source>
    <image_filename>SampleImage.aff</image_filename>
  </source>
```

Figure 1: Sample DFXML output produced by *fiwalk*.

ISO 8601 format, and the name of the resulting image.

The DFXML file produced by *fiwalk* includes entries for each volume, partition, and associated file system (for those partitions that contain file systems). For each file system, a set of *fileobjects* corresponding to all of the files and directories identified is reported. If individual files within a disk image are themselves preservation targets, one can map metadata from the associated *fileobject* entry within a technical metadata section to one's schema of choice. If the disk image itself is a preservation target, the reporting tools developed for BitCurator can aggregate the data into an overview of the image's content. These reports can be generated several ways; as human-readable and editable documents; as part of a METS technical metadata section describing the file system(s) and formats of files identified; and as graphical visualizations.

The approaches described here – preserving individual files extracted from a disk image, and preserving the image itself – are not mutually exclusive, and collecting institutions may wish to employ both methods.

2.3 Forensic analysis metadata

The BitCurator environment incorporates Simson Garfinkel's *bulk extractor* software to identify and report on "features" (specific sequences of characters or bytes within the bitstream) contained within a disk image or live file system. Instances of these features are recorded by scanning modules designed to identify specific patterns in the raw bitstream of the disk image, including those that may correspond to potentially private, personally identifying, and sensitive information [10].

Post-processing of the feature files produced by *bulk extractor* generates a series of text files linking each individual feature at an absolute byte offset into the disk image with a specific file where the feature appears (or indicating that the feature appears in an area currently unallocated by a file system).

Institutions can use the output of *bulk extractor* (and associated processing scripts) to make decisions about processing and potentially redacting disk image content. If documentation of due diligence in identifying sensitive information on a disk is a high priority, or if a repository wants to provide cross-drive descriptors and access points (e.g. all email addresses that appear on the disks within or across collections), the repository can choose to retain some or all of the *bulk extractor* feature reports. This may often not be warranted, as the files are often large (hundreds or thousands of lines) and include sequences of escaped characters in non-ASCII encodings including UTF-8. If the disk image itself is the preservation target, one could opt to generate features reports as needed in the future. In either case, the *events* associated with the production of *bulk extractor* reports can be used to record the process by which curatorial decisions (and subsequent actions) about a disk image are made. In cases of redaction (for example, when private information needs to be removed from a publicly-accessible version of the materials), such documentation can be used to provide a precise account of the nature and number of redacted and their locations.

3. METADATA MAPPING

In order to support preservation and access activities for disk images within archival workflows, we are developing mappings from metadata elements encoded in DFXML to a range of metadata schemas including PREMIS, METS, and EAD. In the following sections, we discuss how this work is supported by specific digital forensics tools incorporated into BitCurator, along with our evolving model for recording digital forensics preservation actions.

3.1 Creating PREMIS metadata for disk images

When working with tools to create and process forensic disk images, the user may wish to treat the disk images solely as intermediate products in identifying, extracting, and repackaging individual file items (which then become uniquely defined archival objects). In other situations, the user may wish to treat the disk image itself as the primary object to be preserved. For the purposes of this work, we focus on the situations in which the disk image itself is the main preservation target. In practice, the two cases are not mutually exclusive. It is often desirable to

retain both the full disk image and extracted copies of files that were stored on the disk.

Forensic tools support the capture and analysis of two types of metadata relevant to the born-digital lifecycle, specifically with respect to ongoing access and preservation. These activities are primary candidates for the creation of PREMIS metadata events.

First, metadata produced by forensic tools includes information about the physical source from which the disk image is extracted, providing important context about the creation environment. This may include manufacturer information, a serial number, and other hardware specifications. This information may be of interest to future users for historical purposes, and may also assist in supporting future access.

Second, this metadata may describe forensic actions performed prior to submission of a disk image to a repository (including analysis and triage tasks), or produced by the use of forensic tools on disk images contained within archival packages.

We have developed a set of PREMIS objects and events associated with extracting and processing disk images from physical media. Each preservation event is linked to a specific software tool that can be executed by a user of BitCurator. The objects, events, and encodings are described in the following sections.

3.1.1 PREMIS object encoding

PREMIS objects capture significant technical properties about digital objects. A disk image extracted from a physical medium can be treated as the instantiation of the preservation object (P-Object1). It is assigned an *objectIdentifier* with a universally unique identifier (UUID) value generated locally The disk image is described at the file level in accordance with the PREMIS data model, which states that files can be read, written, and copied, and have names and formats [17]. At the time of writing, records of forensic disk formats are absent from format registries such as PRONOM, but the most common formats – including the Expert Witness Format and the Advanced Forensic Format – are well documented online and have mature, robust software access libraries.

The PREMIS container *objectCharacteristics* can be used to capture significant technical properties about digital objects. In our mapping, we use the semantic unit *fixity* to record cryptographic hashes including MD5 and SHA1. These hash values are typically verified prior to ingest to ensure the integrity of the disk image, and to avoid inadvertent alteration and detect bitrot during ongoing preservation actions. PREMIS requires that an object's file format be identified either through the use of *formatDesignation* (containing *formatName* and *formatVersion*) or *formatRegistry* (containing *formatRegistryName* and *formatRegistryKey*). A file format registry can be used to validate formats. Selecting either *formatDesignation* or *formatRegistry* is a local implementation decision based on existing resources and workflow. However, the value of *formatDesignation* can be mapped directly using the metadata produced by Guymager, the open source forensic imaging tool incorporated into BitCurator.

We use the semantic unit *creatingApplication* to capture information about the environment in which the disk image is created. The output from Guymager is processed to extract specific technical details about the creation process, including tool version and time of image creation.

Events and relationships in the digital object lifecycle are also recorded using PREMIS. In order to acknowledge that P-Object1 was created by a specific event, one can use *linkingEventIdentifier* to record the link between the preservation event and the created object. Depending on local repository policies, the location of the original physical media can also be described using *storage/Contentlocation*.

If a repository decides to redact sensitive information from P-Object1, they can create a second PREMIS Object (P-Object2). The redaction tool iredact.py creates a new redacted version of a disk image (P-Object2). The redaction tool output records technical details that are used to describe P-Object2, including fixity information and file format types. We have also mapped tool output to *creatingApplication* to capture details about the image creation environment. It will often be advisable to retain P-Object1 for preservation purposes and make P-Object2 available for public access.

P-Object1 and P-Object2 differ in their relationships. P-Object1 is associated with a specific event (disk image capture) but no other PREMIS objects. P-Object2 can be related to P-Object1 using *relationshipType*, with the input value "derived" and *relationshipSubtype* "derived from." The *relationship/ relatedObjectIdentification/ relatedObjectIdentifierType* type can be used to record the UUID of P-Object1 and the *relationship/relatedEventIdentification* to record the UUID of the redaction event.

3.1.2 PREMIS Preservation Event Encoding

Using the information described in the previous section, we are modeling a set of PREMIS events that capture preservation activities performed on disk images. In this section, we describe five of these preservation events: imaging, file system analysis, feature analysis, report generation, and redaction. We describe each event in turn, along with encoding recommendations for integrating the output of the associated BitCurator tool into an existing repository implementation. Technical details that persist across events (such as unique identifiers assigned by local repositories) are described only in Event 1.

Event 1: Imaging

In the *Imaging* event, the disk image is extracted from the original media source. The event records metadata produced by a capture tool such as Guymager in one of the available forensic formats. One can identify the event by assigning it a unique identifier produced by the local repository. One can then describe the eventType as input value "capture" and use the timestamp produced by Guymager to map to *eventDateTime.* The *eventDetail* is used to record specific features of Guymager, including tool version, compilation timestamp, and associated library dependencies. The *eventOutcome* consists of two possible values: "Image created and verified" or "Image creation failed." The format of the newly-created disk image is mapped to *eventOutcomeDetail* (either .e01 or .aff).

Event 2: File System Analysis

The *File System Analysis event* describes the extraction of the file system(s) from the raw or forensically-packed image. The *file system analysis* event incorporates output from the *fiwalk* tool. We describe the *eventType* as "file system analysis." The BitCurator environment parses the XML file produced by *fiwalk* to capture specific details of the event. As an example, *eventDateTime* records the time of file system analysis and

e*ventDetail* stores specific information about fiwalk. The results of the event, either "file system(s) analyzed" or "failed to identify file system(s)" are mapped to *eventOutcome*. Note that for disk event and the resulting preservation object (P-Object2) by using the UUID of P-Object2 in the linkingObjectIdentifierValue.

Semantic unit	Semantic component	Example value(s)	Derived from
eventidentifier	eventidentifierType	UUID	N/A
eventidentifier	eventidentifierValue	8jb50321-6d7b-4291-89ag-a8b0fhc1f276	N/A
eventType	none		
eventDateTime	none	2013-03-29T16:46:13Z	Report.xml -> Start Time
eventDetail	none	version="bulk extractor 1.3.1"	from Report.xml -> Program, Version, SVN_Version, Compiler
eventOutcomeInformation	eventOutcome	report generated; report not generated	
eventOutcomeInformation	eventOutcomeDetail	Log output of reporting tool	
linkingAgentIdentifier	linkingAgentIdentifierType	preservation system	
linkingAgentIdentifier	linkingAgentIdentifierValue	[name of preservation system]	
linkingAgentIdentifier	linkingAgentIdentifierRole		Institution-specific
linkingObjectIdentifier	linkingObjectIdentifierType	UUID	
linkingObjectIdentifier	linkingObjectIdentifierValue	4bc90445-8d7b-8032-23cb-b7a2cah2e358	

Table 1: Sample encoding of a *Feature Analysis* event using *bulk extractor*.

images containing more than one partition, a partial analysis outcome is possible.

Event 3: Feature Analysis

The *Feature Analysis* event describes forensic analysis of the raw bitstream, which identifies features of interest to BitCurator users. This event incorporates those reports output by *bulk extractor*. Similar to Events 1 and 2, the repository assigns an event UUID. The *eventType* input value is "feature analysis." *EventDateTime* and *eventDetail* can be mapped from the <Execution Environment> section of the XML report produced by *bulk extractor*. An example of the relevant PREMIS encodings for the *Feature Analysis* event is provided in Table 1.

Event 4: Report Generation

The *Report Generation* event describes the collation and aggregation of intermediate forensic metadata into actionable, human-readable reports as PDF files or editable .xlsx files that may be used to inform additional preservation actions. The *eventOutcome* specifies the success or failure of report generation, which can include a string or integer representation of "none."

Event 5: Redaction

The *Redaction* event describes the process of eliminating potentially private and sensitive data from a disk image or copy thereof. In the BitCurator environment, the *iredact.py* Python script distributed with Simson Garfinkel's DFXML tools can be used to overwrite specific patterns within a disk image according to a rule set provided by the user. The *EventDateTime* and *eventDetail* are mapped from tool output. The *eventOutcome* specifies either "redaction completed" or "redaction not completed." In this event, *eventOutcomeDetail* records the full name of the newly created disk image, including its file format. One can also create an explicit relationship between the redaction

3.2 Encapsulating descriptive, administrative, and technical metadata for preservation

METS records metadata on acquisition, management, preservation, and access activities. Local METS profiles and tools used to process such metadata can vary significantly in coverage and functionality.

To support interoperability among institutions and existing collections, we are developing metadata export routines that encapsulate Digital Forensics XML produced by software such as *fiwalk*. These include the *fileobject* described in Section 2.2, automatically generated descriptive metadata, and general technical metadata about file systems encountered within disk images.

4. DISCUSSION AND FUTURE WORK

The metadata production and transformation methods described here are intended to supplement and enhance workflows organized around existing archival processing systems. As part of our ongoing work, we are continuously reviewing and responding to feedback from existing users of BitCurator, enhancing the capabilities of the environment, and streamlining tool implementations.

Proposed changes to version 3.0 of the PREMIS data dictionary will enhance lifecycle support for born-digital materials [4]. One proposal involves transforming the semantic unit *Environment* into its own entity (alongside *Object, Event, Agent, and Rights)*, enabling PREMIS to record and capture important metadata about the computing environment. Such enhancements aim to further enable rendering and deployment of preserved digital objects over the long term [5]. These proposed changes complement BitCurator's objectives by providing the necessary structure to

preserve critical metadata describing original software environments.

5. CONCLUSION

We have detailed work conducted as part of the BitCurator project to develop strategies for transforming and incorporating metadata produced by digital forensics tools into preservation and archival metadata schemas. We have shown how metadata produced by open-source digital forensics tools can be encoded into PREMIS events and objects, and then packaged along with other archival metadata in XML format for long-term access and preservation in a repository setting.

6. ACKNOWLEDGMENTS

This research is being performed as part of the BitCurator project, funded by the Andrew W. Mellon Foundation.

7. REFERENCES

[1] AIMS Working Group. "AIMS Born-Digital Collections: An Inter-Institutional Model for Stewardship." 2012.

[2] Cohen, M., Garfinkel, S. L., and Schatz, B., Extending the advanced forensic format to accommodate multiple data sources, logical evidence, arbitrary information and forensic workflow, Proceedings of DFRWS 2009, Montreal, Canada, 2009.

[3] Dappert, A., PREMIS Tutorial: Understanding & Implementing the PREMIS Data Dictionary for Preservation Metadata. Presented at the PREMIS Tutorial, Rome, Italy, 2009. Retrieved 11 June, 2013 from http://www.loc.gov/standards/premic/premis-Rome-pt1.ppt

[4] Dappert, A., Proposed Data Model Changes for PREMIS 3.0, PREMIS Implementation Fair, October 2012. Retrieved 11 June, 2013 From http://www.loc.gov/standards/premis/pifpresentations-2012/PREMIS_Data_Model_Changes_final.pdf

[5] Dappert, A., Peyrard, S., Delve, J., and Chou, C., Describing Digital Object Environments in PREMIS, In *Proceedings of the Ninth International Conference on Digital Preservation (iPRES)*, Toronto, Canada, October 1-5, 2012, pp. 69-76

[6] Encase image file format, http://www.forensicswiki.org/wiki/Encase_image_file_format, Retrieved June 21, 2013.

[7] Garfinkel, S. L., AFF: A New Format for Storing Hard Drive Images, *Communications of the ACM* 49, no. 2, 2006), pg 85-87.

[8] Garfinkel, S. L., Digital Forensics Research: The Next 10 Years, Proceedings of DFRWS 2010, Portland, OR, August 2010

[9] Garfinkel, S. L., Digital Forensics XML and the DFXML Toolset, Digital Investigation 8, 2012, pg. 161-174

[10] Garfinkel, S. L., Digital media triage with bulk data analysis and *bulk_extractor*, Computers and Security, Volume 32, Feb 2013, pp. 56-72

[11] Garfinkel, S. L., "Providing Cryptographic Security and Evidentiary Chain-of-Custody with the Advanced Forensic Format, Library, and Tools." (International Journal of Digital Crime and Forensics 1, no. 1, 2009), pg. 1-28.

[12] Garfinkel, S. L., Lessons Learned Writing Digital Forensics Tools and Managing a 30TB Digital Evidence Corpus, Digital Investigation 9, 2012, pg. S80-S89.

[13] Gengenbach, M. J., "The Way We Do it Here" Mapping Digital Forensics Workflows in Collecting Institutions. Masters Paper for the M.S. in L.S degree. August, 2012.

[14] John, J. L., "Digital Forensics and Preservation", Digital Preservation Coalition, 2012.

[15] Kirschenbaum, M. G., Ovenden, R. and Redwine, G., "Digital Forensics and Born-Digital Content in Cultural Heritage Collections." (Council on Library and Information Resources, Washington, DC, 2010).

[16] Lee, C. A., Kirschenbaum, M. G., Chassanoff, A., Olsen, P., and Woods, K., BitCurator: Tools and Techniques for Digital Forensics in Collecting Institutions, D-Lib Magazine 18, No. 5/6, May/June 2012.

[17] PREMIS Editorial Committee, Data Dictionary for Preservation Metadata: PREMIS Version 2.2. July 2013, p.7.

[18] PREMIS Editorial Committee, Guidelines for using PREMIS with METS for exchange. Library of Congress.

[19] PREMIS Editorial Committee, Use of the Data Dictionary: PREMIS examples. Library of Congress. 2005.

[20] Archivematica, Metadata elements. *Archivematica:* https://www.archivematica.org/wiki/Metadata_elements 2005. Retrieved 13 June, 2013.

[21] Yale University Library Preservation Metadata Task Force, Yale Library, 2006. Retrieved 12 June, 2013 from http://www.library.yale.edu/cataloging/metadata/pmtf/tree.html

[22] Woods, K. and Lee, C. A., Acquisition and Processing of Disk Images to Further Archival Goals, Proceedings of Archiving 2012, Springfield, VA, Society for Imaging Science and Technology, pg. 147-152.

[23] Woods, K., Lee, C. A., and Garfinkel, S. L., Extending Digital Repository Architectures to Support Disk Image Preservation and Access, JCDL '11: Proceeding of the 11th Annual International ACM/IEEE Joint Conference on Digital Libraries, New York, NY, 2011, ACM Press, pg 57-66.

.

Permanent Digital Data Storage: A Materials Approach

Barry M Lunt
Brigham Young University
265 Crabtree Building, BYU
Provo, UT 84602
+1 (801) 422-2264
luntb@byu.edu

Robert Davis
Brigham Young University
N-215 ESC, BYU
Provo, UT 84602
+1 801-422-3238
davis@byu.edu

Douglas Hansen
Millenniata, Inc.
915 S. 500 E. Suite 112
American Fork, UT 84003
+1 801-358-0935
doug@millenniata.com

Matthew R. Linford
Brigham Young University
C-306 BNSN, BYU
Provo, UT 84602
+1 801-422-1699
mrlinford@chem.byu.edu

Hao Wang
Brigham Young University
N-215 ESC, BYU
Provo, UT 84602
+1 801-422-1699
shuaiwanghao1@gmail.com

John Dredge
Millenniata, Inc.
915 S. 500 E. Suite 112
American Fork, UT 84003
+1 801-358-0935
john.dredge@millenniata.com

ABSTRACT

Permanent marks, interpreted as bits, are the *sine qua non* of deep archival storage. Until about 2006, this area of research apparently did not exist, but advances in the past several years at Brigham Young University and at Millenniata, Inc., have produced one product (the M-Disc - a permanent recordable optical disc of DVD capacity), a follow-on recordable optical disc of Blu-ray capacity, work on multi-layer optical discs of Blu-ray capacity on each layer, work on a permanent solid-state storage medium, and work on a permanent optical tape storage medium. This paper explains the approach used to develop these products and advance the research for permanent digital data storage.

INTRODUCTION

For deep archival storage, the ideal is to be able to create the desired artifact, then store it and forget about it, knowing that whenever we wish to access it again, it will still be there, and we will still be able to read or observe it. While this has been the case with non-digital artifacts throughout history, as exemplified by the many historical artifacts remaining from ancient civilizations, this has never been the case with digital artifacts. Starting with ½-inch tape and hard drives, and moving on through SRAM, DRAM, ROM, floppy discs, optical discs, and flash memory, the situation has always been that digital data has been quite ephemeral, especially as compared to historical artifacts. Even the printed book has a lifetime much greater than the best digital storage option, even though books are a technology many centuries old.

In this day of going fully digital, we could greatly benefit from digital storage media which are permanent. While such a development does not solve all the problems of accessibility for centuries to come, it is the *sine qua non* of deep archival storage.

HISTORICAL EXAMPLES

There are many historical examples of artifacts which have survived for centuries and even for millennia. Such examples include coins (metal and ceramic), written documents, vessels, buildings, weapons, clothing, and many types of works of art. Some of these have survived primarily due to the optimal storage conditions in which they were left; a classic example of this would be the Dead Sea Scrolls, found between 1946 and 1956 in caves near the Dead Sea. Lying in the dry climate, with low humidity and little light, and remaining undisturbed for centuries, these documents were remarkably well preserved, especially considering that most of them were made from parchment and papyrus.

Undoubtedly there have been many other documents produced over the centuries which were not stored in ideal conditions, and which deteriorated quickly, leaving us no knowledge of their ever having been created. Others documents were stored in less than ideal conditions, and suffered severe deterioration but were not utterly destroyed.

A materials approach to studying these surviving artifacts reveals much, as would be expected. Artifacts made of gold essentially do not deteriorate; artifacts made of brass, bronze, and silver suffer some deterioration, but have often lasted for millennia. Artifacts made of pottery, ceramic, and other vitreous materials, while brittle, have still not deteriorated much, and remain today as widespread tokens of bygone civilizations.

Using these materials as our examples, it seems obvious that for deep archival storage, artifacts should be made from materials which either do not oxidize (such as gold), or else are by nature fully oxidized or chemically-reacted materials (such as vitreous

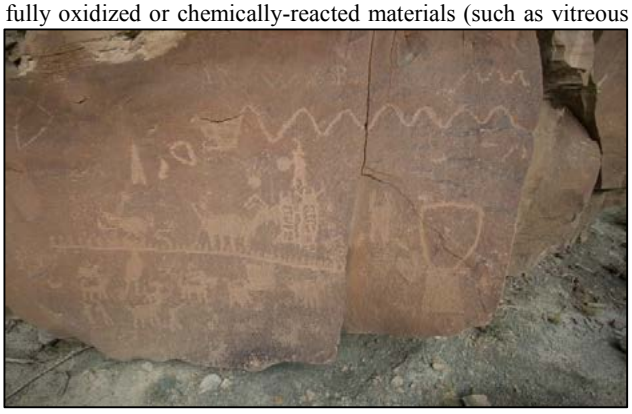


Figure 1: A petroglyph left by the Fremont Indians in Utah, in a place known as 9-Mile Canyon.

Figure 2: A close-up of the petroglyph of Figure 1, showing the etchings in the rock.

materials). Likewise, if we want to preserve writings, the surface on which we write must be have these same characteristics (does not oxidize, or is already fully oxidized), and the ink we use must be treated in such a way that it becomes permanent (such as painting clay then firing it in an oven).

A classic example of an artifact which endures for millennia is petroglyphs (see Figure 1). These writings were made by chipping away a thin layer of dark rock, exposing a lighter layer underneath (see Figure 2). Because these drawings involve physical changes in very durable materials, petroglyphs have endured for centuries in the worst of storage conditions – in the open weather.

NEW TECHNOLOGY, OLD MATERIALS

If we take a materials approach to the problem of permanent digital data storage, it can be seen that if we use extremely durable materials, and if we make significant changes to these materials, these changes will persist, giving us a permanent storage medium.

Accordingly, when this research began in 2006, the first group of materials we studied were metals that either do not oxidize (such as gold), or that form a thin self-limiting layer of oxide (such as aluminum). We knew we could make permanent marks in these metals using relatively cheap solid-state lasers, such as those in CD, DVD, and Blu-ray drives, and we experienced a great deal of success with these materials.

APPLICATION TO CURRENT STORAGE METHODS

Current digital data storage options include magnetic hard-disk drives, magnetic ½-inch tape, flash memory (including USB sticks or "thumb drives", and solid-state drives or SSDs), and optical discs (including CDs, DVDs, and Blu-ray discs). While these storage options have greatly increased in density while dropping in price, they have not addressed the issue of data permanence.

A new technology always has a better chance of success in meeting customer needs if it is not too different from existing technologies. This is particularly true in consumer products for digital data storage. Ac cordingly, we chose to use the existing standard of DVDs, changing only the materials of the recording layer so as to make the data perm anent. This allowed our new DVD-compatible optical disc (the M-Disc) to be widely accessible, as it was readable in all DVD drives. Additional research is currently underway in permanent solid-state storage and in permanent optical tape storage, as described below.

STATE OF THE ART IN PERMANENT OPTICAL STORAGE MEDIA

As anyone experienced in optical data storage will know, merely making a mark in a layer of recording material is not sufficient to produce a robust data storage solution. In our optical disc research, a recording stack was developed consisting of a small number of layers of inorganic, stable materials. Using standard optical disc recording and playback lasers, we were able to reduce the jitter to below the specification for DVDs, and the marks produced (see Figure 3) provided excellent optical contrast and extreme durability.

Optical data recording has a significant advantage when it comes to data longevity, and that is that the recording and playback process does not involve any contact between the media and the recording and playback mechanism. This separation between the recording and playback device and the media allows an infinite number of playback iterations, which is unique when compared to tape. And the relative simplicity of the playback mechanism means that, if the data persists on the media, future optical playback systems will readily be capable of being adapted as necessary to read data stored permanently on optical discs.

A WORKING DEFINITION OF PERMANENT DATA STORAGE

The title of this paper includes the word "permanent". But if permanent means that something endures forever, nothing that science acknowledges is permanent. If we accept a more practical definition, it means that something lasts an extraordinarily long time. But that is a relative term, and lacks a specific meaning.

What would permanent mean, when it comes to data storage? We should probably look to non-digital data storage media for answers. Books are generally considered rather permanent, and if properly manufa ctured and stored, can easily endure for centuries. It would be great if a digital data storage medium could have a similar life expectancy (LE) – a term which should be defined. We accept the definition given by ANSI/AIIM:

"Life Expectancy: Length of time that information is predicted

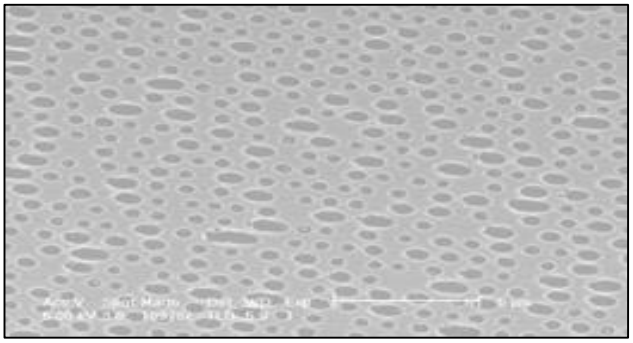


Figure 3: Marks made on the M-Disc, as seen with an SEM.

to be retrievable in a system under e xtended-term storage conditions. Rating for the 'life expectancy' of recording materials and associated retrieval systems. NOTE: The number following the LE symbol is a prediction of the minimum life expectancy in years for which information can be re trieved without significant loss when prope rly stored under extended-term storage conditions, e.g., LE-100 indicates that information can be retrieved for at least 100 years of storage"[5]

One definition for permanence has been proposed for paper:

> "Permanence: The ability of paper to last at least several hundred years without significant deterioration under normal use and storage conditions in libraries and archives."[6]

From a practical perspective, permanence for digital data storage must be defined similarly. We would propose the following:

Permanence: The ability of a digital data storage medium to last at least two hundred years without significant deterioration under normal use and storage conditions in libraries and archives. This means there is a 99.99% confidence of complete data recovery using the intended read mechanism or hardware.

Using the above definition, there is only one digital data storage medium which even comes close to meeting this standard of permanence – the M-Disc, from Millenniata, Inc. All other digital data storage media have very serious limitations when it comes to permanence.[7,8,9,10,11]

The M-Disc is presently available in both DVD-R and BD-R formats, with capacities of 4.7 GB and 25 GB, respectively. Future enhancements include the development of a dual-layer BD-R format and a 2-sided, 2-layer BD-R format, for capacities of 50 GB and 100 GB, respectively.

VERIFICATION OF LIFE EXPECTANCY

One of the first questions raised by the definition we propose is how the LE of a medium can be verified. The science of accelerated testing has long been accepted as a reliable way to determine LE, using the Arrhenius and Eyring equations as the scientific foundation.[12]

For optical discs, the standard method for determining the LE for a given medium has been accepted to be the international standard, ISO/IEC 10995[13]. This standard outlines four test conditions (85°C, 85% RH; 85°C, 70% RH; 65°C, 85% RH; 70°C, 75% RH), as well as the hours required for each set of conditions and the criteria for failure. These tests, even when run in parallel, require a minimum of 2,500 hours of accelerated aging, which is over 104 days at 24 hours/day. Adding in the required testing at appro-priate intervals means that such a test requires many months, much test equipment, and careful testing and analysis. In short, such a test is a major undertaking, and is rarely done in its entirety.

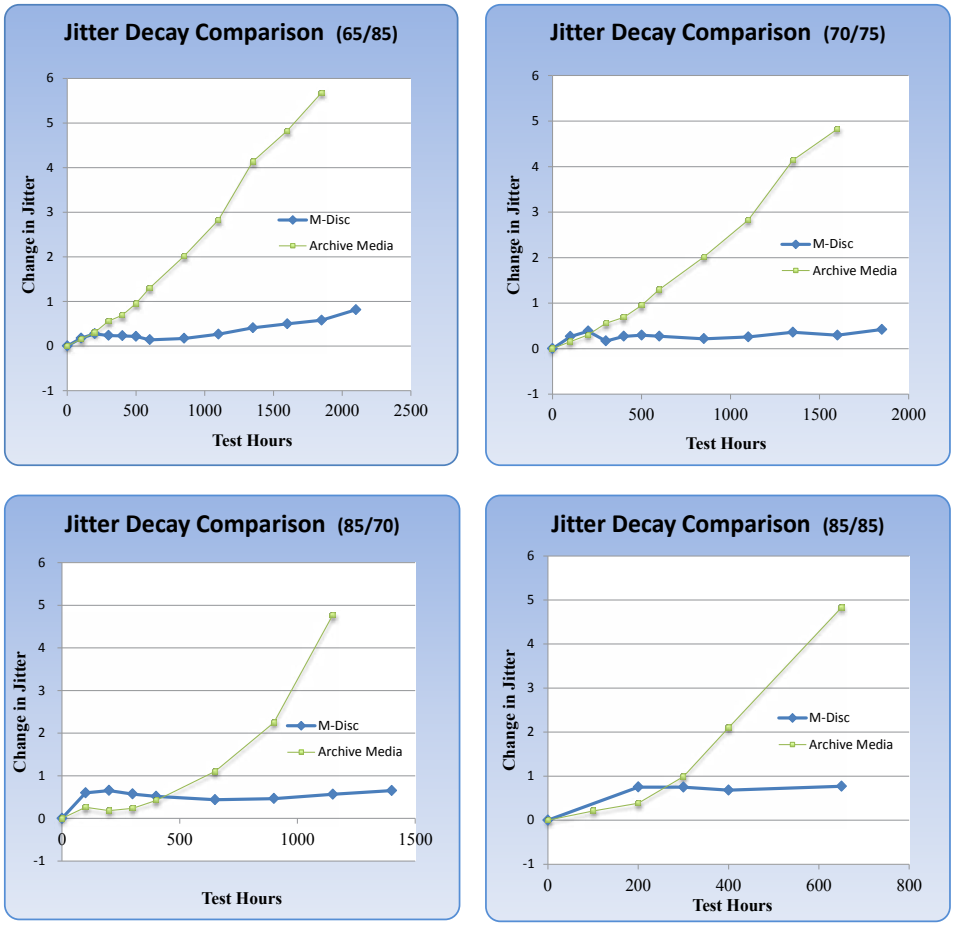


Figure 4: Jitter as a function of test hours, for all 4 test conditions of the ISO/IEC 10995 test.

Nevertheless, a full ISO/IEC 10995 test has been completed on the M-Disc, and the preliminary results have been analyzed. One parameter that has shown a clear difference is jitter, which is an electronic measure of the variation in the shape and position of the physical bits on the disc. If this parameter increases over time, it is a clear indication that something physical is happening to the bits on the disc – a sign of degradation. Figure 4 combines the jitter measurements in all four test conditions specified in the ISO/IEC 10995 test, where the numbers indicate the temperature (in degrees C) and the relative humidity (in percent). The comparison was between the best archive media (JVC Archival Grade DVD-R and Verbatim UltraLife Gold Archival DVD-R) and the M-Disc. In all four test conditions, it is readily evident that the jitter in the archive media is increasing significantly over time, while the jitter in the M-Disc increases only slightly. For example, in the harshest of the four test conditions (85°C/85% RH), over the course of the 600 hours of the test, the jitter in the archive media increased by nearly 5% (far exceeding the specification) and was still climbing, while the jitter in the M-Disc increased by less than 1%, and had apparently stopped increasing. This trend is evident in all four test

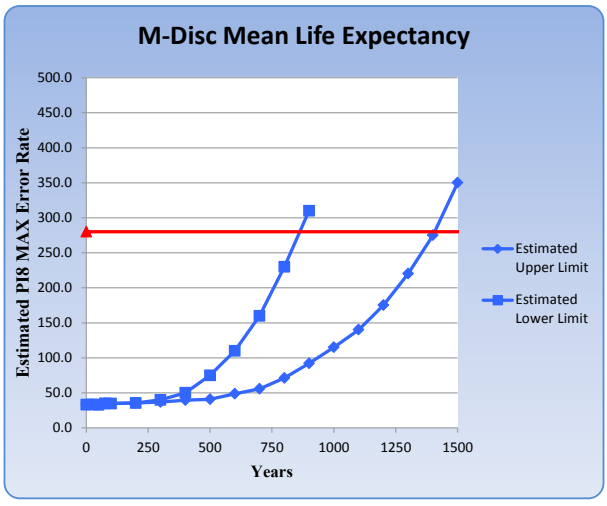


Figure 5: Mean LE for M-Disc, based on ISO/IEC 10995 test results.

conditions, and clearly shows that the best archive media available is not stable with time, while the M-Disc is extremely stable in all four test conditions.

The PI8 maximum error rate is the main specification by which the lifetime expectancy (LE) is estimated. The specification limit is 280. Figure 5 shows the mean LE for the M-Disc, based on the results of the testing that has been done. From these results, it appears that the LE of the M-Disc is between 750 and 1400 years.

Research has been conducted in applying these same recording-layer materials (from the M-Disc) to a Blu-ray type disc. These results have been highly successful, resulting in the announcement at the 2013 Consumer Electronics Show (Las Vegas, NV, Jan 9, 2013) of the availabil ity of a Blu-ray density M-Disc, to be manufactured by Ritek, with availability in July 2013. Lifetime testing has not yet been performed on this disc, but because it uses the same materials approach of the M-Disc and is thus a unique type of archi val disc, it is highly probable that the lifetime will also be dramatically better than anything presently on the market. We also believe that these same materials can be applied to produce a permanent optical tape; this is discussed in more detail later in this paper.

STATE OF THE ART IN PERMANENT SOLID-STATE STORAGE MEDIA

Solid-state storage consists today of DRAM, SRAM, and Flash memory; the characteristics of these memory types are summarized in Table 1. Two of these memory types are volatile, which means that the data disappears if power is lost; these memory types are thus useless for archival storage. Flash memory is the only main type of memory available today that is nonvolatile, which is what has made "flash drives" (also known as "thumb drives") not only possible but also very popular. However, for archival storage they are not practical, since the data is stored as a charge on a floating gate (see Figure 6), and this charge will eventually leak away. Most estimates are that this data will persist for only 8-10 years.

Table 1: Characteristics of main memory types available today

Characteristic	DRAM	SRAM	Flash
Write speed	50 ns	10 ns	500 ns
Read speed	50 ns	10 ns	100 ns
Data retention	Volatile	Volatile	8-10 years
Relative cost	1.0	4.0	1.0
Typical density	16 Gb	4 Gb	16 Gb

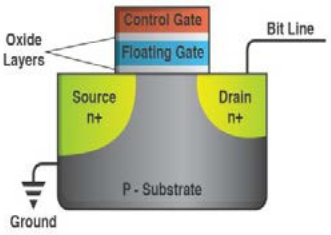


Figure 6: Diagram of a flash memory cell, which stores data as a charge on a "floating" gate.

The materials approach to understanding these solid-state storage media explains why they are not permanent. For DRAM and SRAM, their structure is such that they cannot store data at all, unless they have power present; the "1" state is only distinguishable from the "0" state if power is continuously applied. Flash memory is quite different, requiring power to write and read the data, but not requiring power to maintain the data, at least for several years. However, because it is erasable, there is very little electrical difference between the "1" state and the "0" state, and there is no material difference between these states.

If a storage medium is erasable, by nature it is not as permanent as a medium that cannot be erased. If data is recorded by a physical change in the medium (as in scribing on gold plates; see Figure 7), such data storage is inherently much more permanent that any non-physical change. Another way of saying this is that the greater the difference between the "1" state and the "0" state, the more persistent the data will be.

Figure 7: An Etruscan gold plate with lengthy inscription, dated to about the 5th century BCE.

The early da ys of solid-state storage (the 1960s and 70s) included the development of s everal types of non-volatile memory: ROM (read-only memory), PROM (programmable read-only memory), EPROM (erasable PROM), and EEPROM (electrically-erasable PROM). ROM cannot be programmed by the user, and so is not useful for end users – it must be programmed when it is manufactured. PROM was very useful for niche applications, but it suffered from a defect known as dendrites (see Figure 8) – small tree-like growths that tended to grow and short out the

Figure 8: A dendrite has grown from the remains of a programmed cell (blown fuse) in a PROM storage device.[14]

blown fuse, thus compromising the data. EPROM and EEPROM are the immediate predecessors of today's flash memory, which has been discussed already. Thus, historically, none of the solid-state storage options has been viable for permanent digital data storage.

For solid-state storage to be permanent, both the programmable cells (where the bits are actually stored) and the read/write circuitry must be permanent. Fortunately, half of this problem is already solved, since integrated circuits (ICs) have lifetimes typically rated in FITs – Failures In Time, or the number of devices that fail in 10^9 hours of operation (approximately 1,000,000 years). Typical FIT numbers for ICs are in the range of 50, which means that the typical IC can be expected to operate for over 22,000 years.

One way in which the data in a permanent solid-state memory device could be stored would be to represent a 1 with an intact fuse, and a 0 as a blown fuse. The user could then program their own data into the memory. But the material for the fuse must be as long-lasting as the IC itself. To solve this other half of the problem of permanence for solid-state digital data storage, we need a materials approach – we must find a material which is extremely stable, somewhat resistive but not an insulator, and which does not grow dendrites when the fuse is blown. This would produce fuses which are extremely long-lasting either intact or as programmed. Fortunately, such materials exist, and we have been successful in producing and programming fuses of these materials.

There is still much development work to be done before ICs using these design concepts can be commercially available, but the fundamental concepts have been proven in the lab and in lifetime testing.

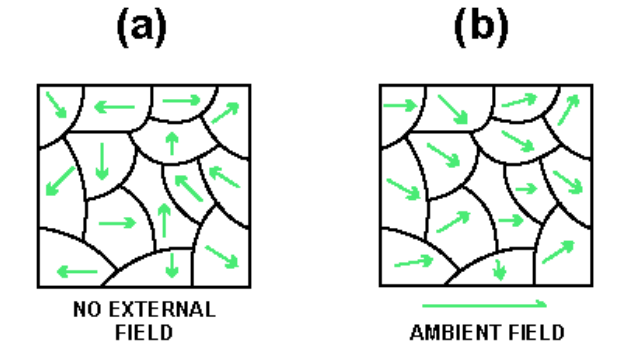


Figure 9: Schematic drawing of magnetic domains, subject to no external magnetic field (a), and subject to an ambient magnetic field (b).

STATE OF THE ART IN PERMANENT ½-INCH TAPE STORAGE MEDIA

Current ½-inch tape products are all magnetic, and suffer two main degradation mechanisms. The first is the slow relaxation of the magnetic domains; these domains are how the data is stored in all magnetic storage products. These domains are depicted in Figure 9. When not under the influence of any external magnetic field (Figure 9a), these domains are randomly oriented. After being subjected to an ambient magnetic field, these magnetic domains change to produce a net magnetization in one of two directions. This difference in the direction of the remaining magnetic field is the difference between 1s and 0s in digital data. With time (and with temperature), these magnetic domains begin to relax, slowly reverting to their original random orientation, and slowly degrading the difference between the encoded 1s and 0s. Eventually, so many of these bits will have degraded that reading a file back will have become impossible.

The other main degradation mechanism of magnetic tape is the delamination of the recording layer from the plastic substrate of the tape. In modern ½-inch magnetic tape, the recording layer is applied to the polyester substrate in a printing process. The binder in the recording material is what keeps the recording material bound to the substrate. But binder materials are generally organic, and they degrade with time, which leads to small pieces of the recording layer peeling off ("delaminating") from the polyester substrate. Wherever these pieces peel off, the data is irretrievably lost.

The two main advantages of magnetic storage are that it is non-volatile, and that it can be re-recorded an infinite number of times. However, as mentioned before, as soon as we make a medium in which the data can be erased (changed), we have lowered the barrier to losing that data with time. We also need to solve the problem of delamination. A materials and processes approach to this problem would be to choose a material which can be readily and permanently altered with a laser in a way that can be detected with a laser, and then to deposit this layer of materials in such a way that they are permanently bonded to the polyester substrate without the use of any binders. Because we have already done this with the recording layer of the M-Disc, our research has focused on applying these materials and processes to a new optical tape, one which is as permanent as the M-Disc.

Figure 10 shows some marks which have been made on this recording layer in ½-inch tape, using less than 10 mW of optical power, which is readily available

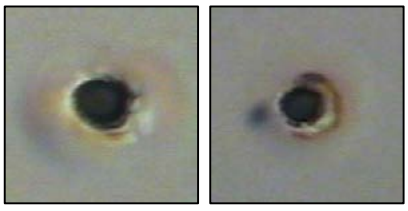


Figure 10: Marks made in our recording layer on ½-inch tape, using less than 10 mW of optical power.

from today's solid-state lasers as used in optical disc drives. These marks are expected to be as permanent as those on the M-Disc – they should endure for centuries.

We have also used a very different process for depositing the recording layer on the polyester substrate, a process which uses no binder, and which is widely used in industry today. We have tested the adhesion of the recording layer to the substrate, using both the tape test method and using a tape tension & bend method. Both of these tests were rigorously applied, and no material was removed. We have shown that the recording layer is permanently bonded to the substrate, so delamination will not occur.

Using a mark density equal to that in Blu-ray discs, and a track density equal to that of today's ½-inch tap e, we would expect that an optical tape cartridge should be able to hold multiple Terabytes, which is roughly equivalent to the density of today's ½-inch magnetic tape.

CONCLUSION

We believe there is an urgent need for some way to store digital data permanently[3,4], and that the M-Disc is one viable solution to this need. We also believe that we are well on our way to providing two additional media for permanent digital data storage.

We readily acknowledge that these permanent media are not a complete solution – there must also exist some way to read the data stored on them far into the future. But data on a permanent medium is the *sine qua non* of deep archival data storage. Some future generation may struggle to learn how our data was stored, but if all the marks on the Rosetta stone had faded away, deciphering them would never have been possible. It is relatively easy and inexpensive to read the marks made on optical discs - the hardware is widely available and optical discs are the most widely adopted digital storage medium in history. And wide adoption is a powerful predictor of relative permanence of readability, as witnessed by the many people who still learn to read and write Latin; though a "dead" language, there are hundreds of thousands of documents in that language. If the marks on these Latin documents were to suddenly disappear, Latin would become irrelevant.

REFERENCES

1. Svrcek, Ivan, "Accelerated Life Cycle Comparison of Millenniata Archival DVD", *Life Cycle and Environmental Engineering Branch, Naval Air Warfare Center Weapons Division*, China Lake, CA, USA, Nov 10, 2009.
2. Lunt, Barry M., Matthew R. Linford, "Towards a True Archival-Quality Optical Disc", *Proceedings of 2009 International Symposium on Optical Memory-Optical Data Storage Conference*, Nagasaki, Japan, Oct 2009.
3. Smit, Eefke, Jeffrey Van der Hoeven, David Giaretta, "Avoiding a Digital Dark Age for data: why publishers should care about digital preservation", *Learned Publishing*, 24: 35-49, Jan 2011.
4. Kuny, Terry, "A Digital Dark Ages? Challenges in the Preservation of Electronic Information", *63rd IFLA (*International Federation of Library Associations and Institutions*),* http://www.ifla.org.sg/IV/ifla63/63kuny1.pdf, September 1997.
5. ANSI/AIIM TR2-1998, "Glossary of Document Technologies", Association for Information and Image Management International, Silver Spring, MD.
6. ANSI/NISO Z39.48-1992 (R1997), "Permanence of Paper for Publications and Documents in Libraries and Archives".
7. Byers, Fred, "Optical Discs for Archiving", Information Technology Laboratory, Information Access Division, NIST, OSTA, Dec 6, 2004.
8. Iraci, Joe, "The Relative Stabilities of Optical Disc Formats", *Restaurator: International Journal for the Preservation of Library and Archival Material*, 2005, 26:2, 134-150.
9. Navale, Vivek, "Predicting the Life Expectancy of Modern Tape and Optical Media", *RLG DigiNews*, Aug 15, 2005, 9:4; also at www.rlg.org/en/page.php?Page_ID=20744#article3
10. Shahani, Chandru J., Basil Manns, Michele Youket, "Longevity of CD Media: Research at the Library of Congress", Preservation Research and Testing Division, Washington, DC.
11. Slattery, O., Lu, R., Zheng, J., Byers, F., Tang, X, "Stability Comparison of Recordable Optical Discs- A study of error rates in harsh conditions," *Journal of Research of the National Institute of Standards and Technology*, 109, 517-524, 2004.
12. Gillen KT, Clough RL, and Wise J, "Extrapolating Accelerated Thermal-Aging Results: A Critical Look at the Arrhenius Method", Albuquerque, NM, Sandia National Laboratories.
13. ISO/IEC 10995: Information technology – Digitally recorded media for information interchange and storage – Test method for the estimation of the archival lifetime of optical media", ISO/IEC 2008, Geneva, Switzerland.
14. Pecht, Pecht, Michael, "Handbook of Electronic Package Design", Marcel Dekker, Inc., New York, NY, 1991, p. 59

Automatic Preservation Watch using Information Extraction on the Web

A case study on semantic extraction of natural language for digital preservation

Luis Faria
KEEP SOLUTIONS
Rua Rosalvo de Almeida, 5
Braga, Portugal
lfaria@keep.pt

Alan Akbik
Technical University of Berlin
Einsteinufer, 17
Berlin, Germany
alan.akbik@tu-berlin.de

Barbara Sierman
National Library of the Netherlands
Prins Willem-Alexanderhof, 5
Den-Haag, Netherlands
barbara.sierman@kb.nl

Marcel Ras
National Library of the Netherlands
Prins Willem-Alexanderhof, 5
Den-Haag, Netherlands
marcel.ras@kb.nl

Miguel Ferreira
KEEP SOLUTIONS
Rua Rosalvo de Almeida, 5
Braga, Portugal
mferreira@keep.pt

José Carlos Ramalho
University of Minho
Braga, Portugal
jcr@di.uminho.pt

ABSTRACT

The ability to recognize when digital content is becoming endangered is essential for maintaining the long-term, continuous and authentic access to digital assets. To achieve this ability, knowledge about aspects of the world that might hinder the preservation of content is needed. However, the processes of gathering, managing and reasoning on knowledge can become manually infeasible when the volume and heterogeneity of content increases, multiplying the aspects to monitor. Automation of these processes is possible [11, 21], but its usefulness is limited by the data it is able to gather. Up to now, automatic digital preservation processes have been restricted to knowledge expressed in a machine understandable language, ignoring a plethora of data expressed in natural language, such as the DPC Technology Watch Reports, which could greatly contribute to the completeness and freshness of data about aspects of the world related to digital preservation.

This paper presents a real case scenario from the National Library of the Netherlands, where the monitoring of publishers and journals is needed. This knowledge is mostly represented in natural language on Web sites of the publishers and, therefore, is difficult to automatically monitor. In this paper, we demonstrate how we use information extraction technologies to find and extract machine readable information on publishers and journals for ingestion into automatic digital preservation watch tools. We show that the results of automatic semantic extraction are a good complement to existing knowledge bases on publishers [9, 20], finding newer and more complete data. We demonstrate the viability of the approach as an alternative or auxiliary method for automatically gathering information on preservation risks in digital content.

Categories and Subject Descriptors

H.3 [**Information Systems**]: Information Storage and Retrieval; H.3.7 [**Information Systems**]: Information Storage and Retrieval—*Digital Libraries*

Keywords

Digital preservation, monitoring, watch, natural language, information extraction

1. INTRODUCTION

Digital assets are continuously endangered by events that threaten user access or even cause irreparable loss of valuable content. These threats belong to many distinct domains, from technological to organizational, economical and political, and can relate to the holder of the content, the producers, the target communities to which the content is primarily destined for, or other internal or external influencers.

In digital preservation, watch (or monitoring) is a key capability that enables the early detection of these threats [5]. As the volume and heterogeneity of assets increases, it becomes infeasible to manually monitor all aspects of the world that may hinder their long-term access. Considering the scale of the problem, the automation of the watch process becomes a necessary step to ensure proper digital preservation.

But to automate the watch process, one needs data on aspects of the world that might afflict the preservation of digital content. This data relates to facts that directly or indirectly represent preservation risks or indicate the need for further analysis. Furthermore, to be able to automatically

reason on the data and infer preservation risks, this data needs to be expressed in machine readable language, i.e. in an explicit and formal specification [12] such as in a controlled vocabulary or an ontology.

Formally specified digital preservation related data can be found in existing repositories like file format registries and tool catalogues, but is commonly incomplete and outdated [24]. Worse, data about other domains that indirectly affect digital preservation, such as organizational or economical, or narrower domains that relate to a specific use case, is hard to find in a formally specified way.

Estimations on the amount of available machine readable data can be found in the Web of Data initiative [7]. Up until September 2011, the Web of Data contained 31 billion RDF triples and about 504 million links from a series of domains like Life sciences, Publications and Media [8]. In contrast, the amount of information contained in the Web is estimated to be orders of magnitude larger; In 2008 Google announced that they have parsed more than one trillion documents[1]. Furthermore, it is estimated that the Deep Web, i.e. the parts of the Web not indexed by search engines, is up to 550 times larger than the Surface Web [6].

However, the size of the Web of Data cannot be directly compared with the whole size of the Web because it is not measured in the same units. We may be able to extract hundreds of relevant RDF statements from the content of one document, but none from another. Nevertheless, for sake of comparison, if we hypothetically assume that, in average, from a document in the Web we could extract 100 RDF statements, than the Web of Data would be 100 thousand times smaller than the Surface Web and 55 million times smaller than the Deep Web.

Information Extraction technologies [23] present a way of making this wealth of unstructured information on the Web available to machine processing. They extract the meaning of information represented in natural language (and other formats) and express it in formally specified data. For preservation watch, the use of Information Extraction technologies has the potential of greatly improving the completeness and freshness of available data and thereby improving the accuracy and completeness of the watch process.

In this paper, we investigate the use of Information Extraction methods in the Web to assist preservation monitoring. We use a state-of-the-art Information Extraction system to gather information for a real case scenario from the National Library of The Netherlands, in which monitoring of scholarly journal publishers is needed. This large scale experiment is conducted as a proof-of-concept to assess the viability and the potential use of such an approach. We derive institutional policies and the requirements for preservation watch for the scenario and identify sources of information with formally specified data, such as manually created registries, for use in the automatic preservation watch processes.

We evaluate the results of the Information Extraction system against information contained in existing registries. The results strongly indicate that the proposed method can be used to automatically gather large amounts of high quality information from large collections of documents on the Web. Furthermore, the results indicate the viability of the proposed approach as an alternative or auxiliary method for automatically gathering information on preservation risks in digital content in order to keep automatic preservation watch more complete and up-to-date.

[1] http://googleblog.blogspot.pt/2008/07/we-knew-web-was-big.html

2. RELATED WORK

In this chapter, we give an overview on preservation policies, automatic preservation watch and the task of Information Extraction.

2.1 Preservation policies

Preservation policies are formulated by an organisation to guide the process of preserving digital information. The SCAPE project[2] is investigating ways of specifying preservation policies in a machine readable format. Based on several guidelines that support the creation of policies, like the OAIS model [14] and the Audit and Certification of Trustworthy Digital Repositories [13] we distinguish three levels of policies:

- the high level or guidance policies of the organization;
- the policies describing the approach the organization intends to take to achieve the goals that are phrased in the high level policies, called preservation procedure policies;
- the lowest level, on which the policies are described in more detail and focused on, for example, special collections and usage, called control policies.

The high level policies and the preservation procedure policies are "human readable". The control policies can be human readable, but need to have a machine readable version in order to be included in automatic preservation processes, such as automatic preservation watch and planning.

2.2 Automatic preservation watch

Automatic preservation watch is a systematic and automatic process of monitoring aspects of the world that might influence the preservation of digital content, revealing preservation risks or opportunities (such as hardware cost reductions). Preservation watch is usually initiated by policies that define constrains or goals that must be monitored from time to time. To be able to automatically watch these aspects of the world, sources of information for various domains must be identified and automatically harvested. This information is then merged into a central knowledge base that ensures information is well structured and normalized. Assessing this information allows one to find preservation risks and opportunities.

Scout: a preservation watch system

[2] http://www.scape-project.eu

Scout[3] is a semi-automatic preservation watch system that provides an ontological knowledge base to centralize all necessary information to detect preservation risks and opportunities [11]. It uses plug-ins to allow easy integration of new sources of information. The knowledge base can be easily browsed and triggers can be installed to automatically notify users of new risks and opportunities. Examples of such notification are: content fails to conform to defined policies, a format became obsolete or new tools able to render your content are available.

Different classes of information sources where identified to be integrated with Scout:

Format registries like PRONOM and UDFR, are online services with structured and formalized information on formats. They have high quality and relevant information for the digital preservation domain but commonly very incomplete. The PRONOM adaptor is already available on the Scout source code. Other generic-domain file format registries exist and are commonly more complete and up-to-date, but have less structured and formalized information.

Software catalogues are online services with information on tools that render, migrate, analyze and compare files of diverse file formats. This catalogues can be digital preservation specific, like the TOTEM[4] and the SCAPE Component Catalogue (under development), or generic-domain software catalogues in which information is not very structured and formalized.

Digital repositories have information on producer activity and the user community access preferences and problems, which certainly concern the repository owner. Also, when aggregated and viewed as a whole, this information can provide insight on the global tendencies and reveal *de facto* standards. A reference implementation adaptor is already available for the RODA repository[5].

Web Archives, like digital repositories, web archives can be of interest for the whole community. The content profile and the renderability problems [17] found with modern browsers can give important insights, and serve as a representative set, of the global internet content and trends. A reference implementation adaptor is already available for the renderability analysis of web archive systems.

Content profiles provide an aggregate view on the content characteristics and metadata, specially of the technical type. When applied to digital repositories and web archives, content profiling can provide precious information needed for preservation. If this information is shared with the community and viewed as a whole, it allows valuable insight on the wider state of curated content and can serve as an up-to-date indicator revealing technology and usage trends.

Experiments with tools, like migrators, assessing their behavior, reliability, completeness and quality are usually made as part of the preservation planning process. The outcome of these experiments can be of much interest to other users that are considering using that tools with similar objectives.

Policies. On-going work on formalizing preservation control policies and their relationship with organization strategies and goals will enable monitoring of some of the organizational objectives and check the repository compliance.

Simulation. Based on the data gathered by Scout, models of digital repositories can be created to predict the consequences of preservation actions [26], which allows the inclusion of forecasts into the knowledge base and be alerted of preservation risks before they actually happen.

Human knowledge. Some of the knowledge needed for digital preservation is still tacit or unstructured. Having humans as a source of information would allow the watch system to act as the central place for any kind of knowledge relevant for digital preservation to be gathered and formalized, even when there is no other specialized platform to support it.

Scout limitations

In Scout, the knowledge gathered from the information sources must be normalized and formally specified, i.e. machine readable. This requirement relates to the need for cross-relating different information and allow automatic reasoning on the knowledge to infer preservation risks and opportunities. Many of the identified information source classes have unstructured information. A possible solution is to require that humans, or crowd-sourcing, to be used to introduce the unstructured information available in the Web. Information extraction technologies could play a important part on automating, or reducing the human effort, of the introduction of unstructured information into the watch system.

2.3 Information extraction

Information Extraction (IE) is the task of extracting structured information from unstructured data such as natural language text. The goal of IE is to make information machine readable. Extracted information is often represented in the form of subject-predicate-object triples, where each triple is the *instance* of one semantic *relation*. The *predicate* indicates the relation to which the triple belongs, while the *subject* and *object* are the two entities between which the relation holds. IE methods use *extractors* that are either manually created or trained using machine learning methods to sieve through large volumes of text and distill for each relation a list of relation instances.

Pattern-based Information Extraction. IE methods often use a pattern matching approach where for each relation a set of *patterns* is defined that if encountered indicate a relation instance. The simplest example of such pattern-based information extraction might be a regular expression that finds e-mail addresses in Web pages. Patterns are often defined as *lexico-syntactic* patterns [1] that match natural language text; We illustrate this with the following example: Suppose we are interested in finding which companies have acquired other companies (such as Google buying Motorola for example). We refer to this common example relation as CompanyAcquisition. As lexico-syntactic pattern we may define "[X] acquired [Y]", where "[X]" and

[3] http://openplanets.github.com/scout/
[4] http://keep-totem.co.uk
[5] http://www.roda-community.org

"[Y]" are placeholders for the subject and the object entities respectively. If this pattern matches a statement in natural language text, a triple for the corresponsing relation is extracted. So, if the extractor encounters the sentence "*Google acquired Motorola in 2011.*", the relation instance CompanyAcquisition(Google, Motorola) [6] is extracted.

Challenges and limitations. As each relation is expressed in natural language text in a multitude of ways, one core challenge of Information Extraction methods is finding all patterns that belong to a specific relation. When aiming to extract relations from an open domain corpus such as the Web, this problem becomes more challenging as there may be a potentially unbounded number of relations, for each of which one extractor with a set of patterns must be defined. When interested only in information from a specific domain of the Web (such as digital content preservation), another challenge is to identify and gather relevant natural language text upon which Information Extraction is performed. Current IE methods are mostly limited to working on *explicit* statements in natural language text; reasoning or inferring knowledge from implicit statements is a topic of current research [16].

Level of supervision. A range of research investigates how to reduce the workload of manually defining patterns with machine learning mechanisms that require different amounts of supervision. Approaches range from supervised [19] or declarative [15] approaches, to weakly supervised [18] and unsupervised methods [4]. Supervised and declarative approaches generally produce high quality extractors, albeit at a cost in human effort, while unsupervised approaches are useful for information discovery.

Our approach. In this paper, we apply both a declarative [2] and an unsupervised [4, 3] approach to address a specific information need in the domain of digital content preservation. Our goal is to find relevant information on the Web. We discuss our system and how we address the above stated Information Extraction challenges in Section 3.

3. CASE STUDY

In this chapter, we describe a specific real world scenario in which large amounts of machine readable knowledge are needed. This scenario represents a use case that requires information to be as *complete* and as *up-to-date* as possible, and illustrates how such information can be found in natural language statements on the Web. We derive semantic relations from this use case and use a state-of-the-art Information Extraction pipeline to gather data from the Web and extract the required information from natural language text. We describe each step of the Information Extraction process and perform an analysis of the extraction results.

The purpose of this experiment is to execute a *proof-of-concept* on the idea of using Information Extraction methods to assist preservation monitoring. In particular, we wish to examine the following questions: What is the potential of using IE technologies to assist preservation monitoring? What are limitations and challenges? What are the prospects for future work in this field?

[6]Often, relation instances are denoted by first giving the predicate (i.e. relationship type) in camel case, and then the subject and object entities in brackets separated by a comma.

3.1 A real world scenario

As scholars have become increasingly reliant on electronic versions of scholarly journals, long-term preservation of these resources has become of major importance and a growing need for the library community. The shift to journal content that is digital, online and held remotely has challenged the responsibility that libraries have in ensuring the continuity of access to these materials. The National Library of The Netherlands (KB) was one of the very first cultural heritage institutions to become aware of the emerging importance of digital resources. As early as 1998 the KB concluded an agreement with the Dutch Publishers Association to extend the Dutch voluntary deposit scheme to off-line electronic publications, and in 1999 a tender was issued for the development of a long-term storage facility for electronic information resources. As no ready-made commercial products were available at the time, the KB embarked on a joint project with IBM to develop the Digital Information Archiving System (DIAS). With the establishment of the e-Depot the KB has created in 2002 the first solution to provide permanent access to scholarly information. This goes beyond the national depository role of the KB as it also preserves publications from international, academic publishers that do not have a clear country of origin. Originally, the e-Depot was designed to preserve the electronic publications of the Dutch publishers, in agreement with the Dutch voluntary deposit scheme. Some of the early archiving agreements were signed with major scientific publishers based in the Netherlands, such as Elsevier and Kluwer. As these are internationally operating publishers, the question soon arose how digital resources which are simultaneously published all over the world, fit into traditional national deposit schemes. The answer was simple: they do not. The KB decided that a new international framework would have to be developed to preserve digital publications for the long-term. As such a framework does not come to be overnight, the KB took a step by opening up its own e-Depot facilities to digital resources published by international publishers. Content for the e-Depot is delivered directly by scholarly publishers who have agreed to participate in the KB archiving service. As of June 2012, the e-Depot has preserved over 18 million journal articles.

The problem with e-journals

Today there are three leading archiving organizations agreed to act as last resort for e-journal content. Besides KB e-Depot, Portico[7] and CLOCKSS [8] are providing permanent access to this type of digital materials. All three are working very closely together and are involved in the Keepers registry which is a resource to address "who is looking after which e-journals, how, and what are the terms of access?"[9].

The next step for the KB is to position the international e-Depot as a European service, which guarantees permanent access to international, academic publications on a European level [22]. There is a danger that e-journals become

[7]`http://www.portico.org`
[8]`http://www.clockss.org`
[9]`http://www.thekeepers.org`

Table 1: Distribution of titles per publisher

% of titles	Publishers	Titles per publisher
40%	9	> 310
50%	21	> 132
60%	52	> 52
70%	143	> 16
75%	267	> 7
80%	569	> 3

Table 2: Publisher (P.) size distribution [10]

P. Size	# journals	% of P.	% of articles
very small	1-10	97%	30.9%
small	11-50	2%	14.6%
medium	51-250	0.32%	6.9%
large	250-1000	0.04%	6.2%
very large	> 1000	0.08%	41.4%
Total	**17.565**	**4.993**	**1.628.354**

"ephemeral" unless we take active steps to preserve the bits and bytes that increasingly represent our collective knowledge. Besides the threat of technical obsolescence there is the changing role of libraries. In the past, libraries have assumed preservation responsibility for materials they collect, while publishers have supplied the materials libraries need.

These well understood divisions of labour do not work in a digital environment and especially so when dealing with e-journals. So we need new models and organizations to ensure safe custody of these digital objects for future generations. The KB has invested in order to take its place within the research infrastructure at European level and the international e-Depot serves as a trustworthy digital archive for scholarly information for the European research community.

The scalability problem

To preserve scientific publications for future research we need to keep as much as possible. That means that the e-Depot needs to cover as much as e-journal titles as possible.

According to Ulrichsweb[10], there are over 35.000 peer reviewed journal titles within the academic realm. Over 65% of them, about 23.000, are online journals. According to EBSCO[11] there are over 5.000 publishers who are publishing 25.000 electronic journals. Yet another number comes from Web of Science[12]. This gives over 12.000 e-journals from 3.200 publishers. Looking more closely to these numbers we find out that the 100 largest publishing companies publish almost 70% of the available titles, as shown in the Table 1. So 80% of the available journal are provided by 569 publishers. Beyond that we find a huge long tail. According the the numbers of EBSCO again there are 466 publishers with two journals and 4.000 publishers with only one journal. A similar view is given by Scopus[13], the citation-index of Elsevier. In 2009 it counted almost 5.000 journal publishers in its database. 97% of them publishes 1-10 journals. This is, however, a significant part of the available journal articles, over 30%. In other words, the long tail is very large and in this we have to deal with a large amount of companies, as it is shown in Table 2.

The Tables 1 and 2 show that there is a great deal of concentration of journal titles with a small group of publishers. With 21 large publishers we cover 50% of the journal titles listed by EBSCO. But they also show that we have to face a huge long tail with 80% of the publishing companies publishing only one title. For the coverage of an e-journal archiving service like the KB e-Depot it is fairly doable to sign agreements with the largest publishing companies and ingest their content in the archive. But after that the real work begins, knowing also that each year over 1.5 million scientific articles are published.

Coverage

The international e-Depot was set up to be a service for the European research community to give access to scientific e-journals in case the university repositories or the publishers' platforms, which currently provide access, are no longer available or able to do this. The coverage of the journal titles to be archived is of most importance. Archival services have the aim to cover as many titles and articles as possible. Collections need to be complete. In practice, many situations can influence this completeness, like publishers getting out of business or journals changes between publishers. This happens very often and is a real problem, not only for archives, but certainly also for libraries, who are the subscription payers. The transfer of a title from one publisher to another itself is not the problem. The problem is in the administration of the transfer. Users, like libraries and archives, need to know when a title has been transferred and which publisher has taken over the title and under which conditions. The Transfer Code of Practice from the UK Serials Group gives a set of rules for transferring journal titles:

> The Transfer Code of Practice responds to the expressed needs of the scholarly journal community for consistent guidelines to help publishers ensure that journal content remains easily accessible by librarians and readers when there is a transfer between parties, and to ensure that the transfer process occurs with minimum disruption. The Code contains best practice guidelines for both the Transferring Publisher and the Receiving Publisher. Publishers are asked to endorse the Code, and to abide by its principles wherever it is commercially reasonable to do so. [25]

So the code exists to facilitate the users but, in the real world, this does not always work. Publishers do not follow these rules or do so very late. Administrative handling has no priority for a publisher and is only done months after the actual transfer. This is very problematic not only for the libraries using the subscription, but also for the archives who expect titles to be received from publishers. But after

[10] http://www.ulrichsweb.com
[11] http://www.ebsco.com
[12] http://wokinfo.com
[13] http://www.scopus.com

a transfer it suddenly ceases to receive the title any more. This hinders the coverage and completeness of the archive. It also brings along a great deal of work in finding out where the title has gone and who is the new publisher. So it takes time and work and it is a problem for coverage.

3.2 From the scenario to information sources

If we translate this description of the International e-Depot into polices, we can see that the high level aim is to create a complete collection of international scholarly e-journals for long-term preservation and access by acquiring these e-journals from the publishers in order to serve the European research community, in case the university repositories are no longer able to do this.

In order to achieve this, various preservation procedure policies need to be developed. The list of scholarly e-journals needs to be identified and the related publishers need to be contacted. Once the relationship is established via an agreement, regular monitoring needs to take place in order to be assured that changes can be dealt with and that the goal of "completeness of the journal collection" will be achieved. For this monitoring, detailed control policies will need to be established. The following list describes indications of the situations that can occur:

- Publisher A had journal J, is there a journal J provided by publisher A at time T_1?
- Publisher A had journal J_1 and the journal has been renamed to J_2 (i.e. has changed title or ISSN), is there a journal J_2 provided by publisher A at time T_1?
- Publisher A transferred journal J to publisher B, is there a journal J provided by publisher B at time T_1?
- Journal J has ceased to exist, what is its most recent issue?

In order to monitor these situations, the search results need to be filtered automatically, based on control policies. This experiment is limited to the investigation whether it is feasible to acquire relevant information from the Web and relevant registries. This relevant information relates to existing scientific journals, identified by title or ISSN, and journal-publisher relations that specify which publishers provide a certain journal. This information is manually maintained by registries within the e-Depot and also in other similar repositories and it is aggregated in the Keepers registry. But the information in the registries is only relative to the journals they collectively keep, being difficult to use to it to ascertain the completeness of the journal safeguarding. Furthermore, due to the manual processes involved and the lack of cooperation from the publishers, is incomplete and outdated. Nevertheless, publishers provide this information on their Web sites in natural language. So there is a possibility here for expanding and improving the available information by using information extraction technologies. In the following experiment we will focus on using information extraction technologies to gather information that would allows us to detect the first situation described above, whereas a journal J is provided by publisher A at time T_1.

3.3 Experiment

In our experiment, we aim to find a list of journal titles discussed in the Web, as well as a list of journal-publisher attributions in order to discover which publisher publishes which journal. The experiment consists of three steps; First we execute a *data acquisition and pre-processing step* that first gathers relevant natural language data from the Web. We then perform *relation discovery* on this data to mine frequent extraction patterns and gain an insight into the semantic content of the crawled corpus. We then assign patterns to the relations we wish to mine from the corpus and execute a *relation extraction* step to mine instances for each relation of interest.

Data Aquisition and Pre-Processing

The first phase of the experiment is a data acquisition and pre-processing pipeline. Its goal is to gather large amounts of natural language text from the Web that has a high probability of containing statements that relate to our use case.

We implemented a focused crawler to address this task. It uses a list of seed keywords (such as publisher names and journal titles) and formulates a search query using a Web search engine API[14] for each keyword on the list. Each query returns a list of Web pages that is automatically crawled and processed with Natural Language Processing (NLP) tools. Boilerplating is applied to extract blocks of natural language text from each Web page, removing other Web page elements such as layout information or advertisements. Sentence segmentation is applied to divide blocks of text up into sentences that can be analyzed individually. Finally, we filter out all sentences that do not contain at least one of the seed keywords.

The resulting dataset consists of approximately 18 million sentences gathered from 500.000 Web sites. The total text size is 8 GB. The seed keyword list consists of 12.000 entries. A sample of seed keywords and gathered sentences is illustrated in Table 3. These example sentences contain information relevant to our domain.

Relation Discovery

In the second phase of the experiment, we are interested to discover what kind of relations are expressed in the dataset. Because manually going through a dataset of this size is infeasible, we apply a relation discovery mechanism to identify prominent extraction *patterns* in the text. We apply a method explained in detail in [4] that counts and groups patterns according to distributional evidence in the corpus.

This yields a list of prominent patterns in the corpus, a sample of which is given in Table 4. These patterns indicate that the dataset gathered by the focused crawler is indeed relevant to our domain and suggests relationship types for which extractors can be created. Note that the patterns we use are actually more complicated lexico-syntactic patterns, but the syntactic elements (which denote grammatical properties of the patterns) are not indicated for the sake of readability.

Information Extraction

[14]In our pipeline, we use the Bing API, available at `http://www.bing.com/developers/` (last checked at 2013-04-20)

Table 3: Sample data from the data acquisition and pre-processing pipeline.

Seed keyword	Sample sentence retrieved from the Web
Elsevier	"*In 1991, two years before the merger with Reed, Elsevier acquired Pergamon Press in the UK.*"
The Asia-Europe Foundation	"*The Asia-Europe Foundation (ASEF) sold the Asia Europe Journal and transferred the copyright to its long-time partner Springer.*"
Acta Chirurgica Iugoslavica	"*Acta Chirurgica Iugoslavica is available free of charge as an Open Access journal on the Internet.*"
American Journal of Preventive Medicine	"*The American Journal of Preventive Medicine is the official journal of the American College of Preventive Medicine and the Association for Prevention Teaching and Research.*"
Journal of Business Ethics	"*In 2004 the Journal of Business Ethics merged with the International Journal of Value-Based Management and Teaching Business Ethics.*"

Table 4: Top pattern in the gathered corpus.

Pattern	Rank #
[X] journal of [Y]	1
[X] published by [Y]	2
[X] journal on [Y]	3
[X] journal published by [Y]	4
[X] available as [Y] journal	5
PubMed [X] [Y]	9
[X] science proceedings of [Y]	25
[X] subscription available to [Y]	30

In the third phase of the experiment, we wish to create extractors to find two relations in the crawled document collection: an extractor for journal titles (IsJournal) and an extractor for journal-publisher attributions (JournalPublisher). For each extractor, we manually go through the top patterns found in the relation discovery step and select patterns to use for relation extraction. For the JournalPublisher for example, we assign among others the patterns "[X] journal of [Y]" and "[X] journal published by [Y]".

We then execute both extractors on the document collection and store all found relation instances in two lists: One list of all journal titles found in the Web crawl, and one list of all identified journal-publisher attributions.

Information insertion into Scout

The resulting lists of journal titles and journal-publisher attributions conform to the formally specified and normalized information source restriction of Scout, the automatic preservation monitoring system explained in Section 2.2. This information can be inserted into Scout via a new plug-in, allowing this information to be included into the central knowledge base. Queries and notification triggers can then be created using the information on the knowledge base to alert when journals change publishers, or even to cross-relate an institution's list of subscribed publishers and journals of interest to alert when a journal of interest is no longer provided by any of the subscribed publishers.

The process of finding new journals and journal-publisher attributions used in this experience can be frequently repeated to allow automatic constant monitoring of these aspects of the world, automatically notifying interested users when the preservation risk of not acquiring a journal becomes relevant.

3.4 Results

In the experiment, we generate a list of 2,000 journal titles and a list of 500 journal-publisher attributions. We evaluated the results both automatically and manually against the e-Depot publishers. In the automatic evaluation, we matched the results against the e-Depot to find out how many of the extracted titles were already contained in the e-Depot internal registry[15]. Of the 2,000 journal titles, we found that only 200 were in the e-Depot, making the remaining 1,800 titles candidates for inclusion. We manually went through a sample of 200 of these titles and found that 191 are titles that should be added to the registry.

We manually repeated this experiment with the more complete Keepers Registry and found that more than 50% of all journal titles and 50% of all attributions were not in the registry and should be added. Again we found that the largest part of relation instances were viable candidates for entry into the registry. This indicates a strong potential of using Information Extraction technologies to help in keeping such registries complete and thus aiding the task of preservation monitoring.

In Table 5, we illustrate example instances of the JournalPublisher relation. The sample was chosen by sorting the list of all instances alphabetically by journal title and selecting the first 17 instances. The table illustrates which of these instances is already listed in Keepers Registry and which should be added to make it more complete. Some entries in the list have comments to illustrate error classes such as encoding errors or entity name boundary detection errors.

The information above can be directly used to answer the first situation described in section 3.2, whereas a journal J is provided by publisher A at time T_1, which is the time of data acquisition. The same IE pipeline can be frequently executed to get new snapshots in time, providing a continuous monitoring of this situation. Automatic monitoring of the continuity of e-journal availability can be done by cross-referencing this information about journal-publisher relations with the list of e-Depot paid publisher subscriptions (throughout time) and the list of e-journals available in the e-Depot repository. Nevertheless, for the other situations in section 3.2, more information about the journals needs

[15]the e-Depot archiving service contains an internal registry with the journal titles it archives and its related publishers, this internal registry is aggregated by the Keepers registry

Table 5: Sample of results and comparison to Keepers Registry.

Extracted relation instances		Evaluation	
Journal	**Publisher**	**In Keepers Registry?**	**Comment**
A Journal of Human Environment	Royal Swedish Academy of Sciences	**no**, should be added	
AAPS Journal	Springer Science + Business Media LLC	**yes**	
AAPS Journal	American Association of Pharmaceutical Scientists	**no**, should be added	
Academic Emergency Medicine	Society	**no**, error should be corrected and instance added	Error in entity detection of publisher name. It should be: "*Society of Academic Emergency Medicine*"
ACEEE International Journal of Network Security	ACEEE-Network Security Group	**no**, should be added	
Acta Applicandae Mathematicae	Springer	**yes**	
Acta Automatica Sinica	Chinese Association of Automation and Institute of Automation	**no**	"*Acta Automatica Sinica*" is listed only as published by "*Elsevier*".
ACTA AUTOMATICA SINICA	Chinese Association of Automation and Institute of Automation	**no**	All caps duplicate of previous relation
Acta Biomaterialia	Elsevier	**yes**	
Acta Geologica Slovaca	Comenius University in Bratislava	**yes**	
Acta Materialia	Elsevier	**yes**	
Acta Polytechnica Hungarica	âŤIJĂşâŢęĂěâŤijbuda University	**no**, error should be corrected and instance added	encoding error
Acta Radiologica	Scandinavian Society of Radiology	**no**, should be added	"*Acta Radiologica*" is listed with other publishers.
Aequationes Mathematicae	Birkh	**yes**	Encoding error in publisher name. Should be "*Birkhäuser Verlag*"
African Journal of Biomedical Research	Biomedical Communications Group	**no**, should be added	
Agricultural Economics	IAAE	**no**	"*Agricultural Economics*" is listed with other publishers. "*IAAE*" is an abbreviation for a missing publisher on the list.
Agronomy Journal	American Society of Agronomy	**yes**	

to be captured, like the journal renaming or ceasing. Also, machine readable information on the publisher subscription and e-journal issues available in the repository needs to be inserted into Scout in order to automatically cross-reference and discover other entailed preservation risks. These steps are some of the future work to be done in the Scape project to further study the use of information extraction technologies on digital preservation processes.

3.5 Lessons Learned

The experiment, intended as a proof-of-concept, strongly indicates the viability of using IE methods in preservation monitoring. We proceed to analyze the results more thoroughly with regards to sources of error (see Table 5) and potential improvements.

One major problem that affects extraction quality (i.e. the portion of results that are fully correct) is the lengthy nature of some journal titles or publisher names. An example of this is the "*European Journal of Nuclear Medicine and Molecular Imaging*". This causes our method to detect the wrong title boundaries in some cases; titles might be too short or too long, encompassing either only a portion of the words of the real title, or additional words that do not belong to it. We have adapted our method to cope with this, but more fine-tuning our extractors to this specific domain will arguably increase overall extraction quality.

We make another observation when we revisit the list of patterns in Table 4: We only used a small portion of the top patterns in our experiment. Incorporating additional patterns may lead to more complete extraction results. More importantly, we found that there were many types of information in the crawled corpora that were not extracted but may also be of interest to the community. For example, the pattern PubMed [X] [Y] indicates that information on PubMed entries is contained in the corpus. Similarly, the pattern [X] journal on [Y] indicates that it is possible to extract topics for journals. Accordingly, this indicates potential for expanding the range of information we extract in future experiments.

This experiment shows that the information extraction technologies has potential not only for detection of real-time threats for digital preservation domain, but also for parsing historical knowledge to capture descriptive information and becoming an important tool for librarians and archivists to cope with the increasing scale of digital content production.

4. CONCLUSIONS

Automatic preservation watch becomes a necessary capability of an institution when the factors that must be taken into consideration to do effective digital preservation become too complex or onerous for manual procedures. But automatic monitoring is highly dependent on the available machine readable information about the aspects of the world to monitor. Information extraction technologies can be used to surpass this limitation, allowing the use of information from the Web available in natural language.

The presented case study demonstrates how automatic monitoring can be done by using natural language statements from the Web. A real world scenario from the National Library of The Netherlands is presented where there is a need to monitor the scientific journal publishers, in order to ensure that there is an high coverage of all international scientific journals published throughout the world.

Sources for this kind of information are identified, like the Keepers registry and the e-Depot internal registry, but there are concerns that these registries may be incomplete and outdated. Information extraction technologies are then used to fetch natural language information dispersed throughout the Web and extract journal and journal-publisher attributions automatically. Comparing the information with the Keepers registry we find that more that 50% of the automatically fetched data is not on the registry and should be added, proving that this method is effective and can provide a much needed contribution for the automatic watch of the publisher community.

The technologies and methods used in the use case are not specific to publishing domain and can be applied to other monitoring needs, opening new possibilities for institutions to automate their watch processes. Using information extraction with automated preservation watch systems allows monitoring of non-technical domains, such as social, economical or organizational, where formally specified data is scarce. For example, monitoring economical or organizational changes in companies that support file formats or tools, like company bankruptcy or takeover, may allow the discovery of significant preservation risks. Also, this method allows monitoring of institutional specific domains, like the producer or target community, from which pre-existing formally specified data is rare and mostly manually created by institution itself. Further research on how to use these technologies and methods to monitor digital preservation related domains will be done in the next year of the SCAPE project.

5. ACKNOWLEDGEMENTS

Part of this work was supported by the European Union in the 7th Framework Program, IST, through the SCAPE project, Contract 270137.

6. REFERENCES

[1] A. Akbik and J. Bross. Wanderlust: Extracting semantic relations from natural language text using dependency grammar patterns. In *Workshop on Semantic Search in Conjunction with the 18th Int. World Wide Web Conference*, 2009.

[2] A. Akbik, O. Konomi, and M. Melnikov. Propminer: A workflow for interactive information extraction and exploration using dependency trees. In *ACL System Demonstrations*. Association for Computational Linguistics, 2013.

[3] A. Akbik and A. Löser. Kraken: N-ary facts in open information extraction. In *AKBC-WEKEX*, pages 52–56. Association for Computational Linguistics, 2012.

[4] A. Akbik, L. Visengeriyeva, P. Herger, H. Hemsen, and A. Löser. Unsupervised discovery of relations and discriminative extraction patterns. In *Proceedings of the 24th International Conference on Computational Linguistics*, 2012.

[5] G. Antunes, J. Barateiro, C. Becker, J. Borbinha, D. Proença, and R. Vieira. Shaman reference architecture (version 3.0). Technical report, SHAMAN Project, 2011.

[6] M. Bergman. The deep web: Surfacing hidden value. *The journal of electronic publishing*, 7, 2001.

[7] C. Bizer, T. Heath, K. Idehen, and T. Berners-Lee. Linked data on the web (ldow2008). In *Proceedings of the 17th international conference on World Wide Web*, WWW '08, pages 1265–1266, New York, NY, USA, 2008. ACM.

[8] C. Bizer, A. Jentzsch, and R. Cyganiak. State of the lod cloud. `http://lod-cloud.net/state/2011-09-19/`, 2011.

[9] P. Burnhill. Tales from the keepers registry: Serial issues about archiving & the web. *Serials Review*, 39(1):3–20, 2013.

[10] Elsevier. Scopus. `http://www.scopus.com`, 2009.

[11] L. Faria, P. Petrov, K. Duretec, C. Becker, M. Ferreira, and J. C. Ramalho. Design an architecture of a novel preservation watch system. In *International Conference on Asia-Pacific Digital Libraries (ICADL)*. Springer, 2012.

[12] T. Gruber. A translation approach to portable ontology specifications. *Knowledge Acquisition*, 2(5):199–220, 1993.

[13] ISO. Space data and information transfer systems - Audit and certification of trustworthy digital repositories. ISO 16363:2012, International

Organization for Standardization, Geneva, Switzerland, 2012.

[14] ISO. Space data and information transfer systems - open archival information system (oais) - reference model. ISO 14721:2012, International Organization for Standardization, Geneva, Switzerland, 2012.

[15] R. Krishnamurthy, Y. Li, S. Raghavan, F. Reiss, S. Vaithyanathan, and H. Zhu. Systemt: a system for declarative information extraction. *ACM SIGMOD Record*, 37(4):7–13, 2009.

[16] N. Lao, T. Mitchell, and W. W. Cohen. Random walk inference and learning in a large scale knowledge base. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 529–539, Edinburgh, Scotland, UK., July 2011. Association for Computational Linguistics.

[17] M. T. Law, N. Thome, S. Gançarski, and M. Cord. Structural and visual comparisons for web page archiving. pages 117–120. Proceedings of the 2012 ACM symposium on Document engineering, 2012.

[18] M. Mintz, S. Bills, R. Snow, and D. Jurafsky. Distant supervision for relation extraction without labeled data. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2-Volume 2*, pages 1003–1011. Association for Computational Linguistics, 2009.

[19] R. Mooney. Relational learning of pattern-match rules for information extraction. In *Proceedings of the Sixteenth National Conference on Artificial Intelligence*, pages 328–334, 1999.

[20] E. Oltmans and H. van Wijngaarden. Digital preservation in practice: the -depot at the koninklijke bibliotheek. *VINE*, 34:21–26, 2004.

[21] D. Pearson. AONS II: continuing the trend towards preservation software 'Nirvana'. In *Proc. of IPRES 2007*, 2007.

[22] M. Ras. The international e-depot to guarantee permanent access to scholarly publications. *Cultural Heritage On Line - Trusted Digital Repositories & Trusted Professionals*, 2012.

[23] S. Soderland. Learning information extraction rules for semi-structured and free text. *Machine learning*, 34(1-3):233–272, 1999.

[24] D. Tarrant, S. Hitchcock, and L. Carr. Where the semantic web and web 2.0 meet format risk management: P2 registry. *The International Journal of Digital Curation*, 1(6):165–182, June 2011.

[25] UKSG. Transfer. `http://www.uksg.org/transfer`.

[26] C. Weihs and A. Rauber. Simulating the effect of preservation actions on repository evolution. In *Proc. of iPRES 2011*, pages 62–69, Singapore, 2011.

Preservation Policy Levels in SCAPE

Barbara Sierman
KB National Library of the Netherlands
PO Box 90407
2509 LK The Hague
+31 70 314 01 09
Barbara.Sierman@KB.nl

Catherine Jones
Science and Technology Facilities Council
Harwell Oxford, Didcot OX11 0QX
+44 1235 445402
Catherine.jones@stfc.ac.uk

Sean Bechhofer
University of Manchester
Kilburn Building, Oxford Road
Manchester M13 9PL
+44 161 274 6282
sean.bechhofer@manchester.ac.uk

Gry Elstrøm
State and University Library Denmark
Victor Albecks Vej 1
8000 Aarhus C
+45 8946 2314
gve@statsbiblioteket.dk

ABSTRACT
This paper describes the Preservation Policy model as designed in the European project SCAPE and an experiment to test the viability of the model against two real life preservation policies.

Categories and Subject Descriptors

H.3.7 [**Information Systems**]: Information Storage and Retrieval – *Digital Libraries; I.2.4* [**Computing Methodologies**]*:*Artificial Intelligence – *Knowledge Representation Formalisms and Methods*

Keywords

Digital preservation, policies, watch, planning.

1. INTRODUCTION
There is a shared recognition that the existence of preservation policies for long term digital preservation is important. Not only because it is for example stated in the ISO standard 16363 Audit and Certification of Trustworthy Digital Repositories, but also because digital preservation needs a well defined underlying basis. The creation of these policies seems to be rather difficult and we see that organizations are struggling to write them. Many organizations who are preserving collections for the long term have not yet published their policy on their website. While these organizations often have a legal mandate and are funded by public money, the general public does not know how these digital collections are treated. Nor can they see how these organizations plan to handle various challenges.

A preservation policy is a "Written statement authorized by the repository management that describes the approach to be taken by the repository for the preservation of objects accessioned into the repository".[1]

Preservation Policies are not a goal in itself, they are there to support the activities of the organisation with respect to the maintenance and preservation of the digital collection. "Without a policy framework a digital library is little more than a container for content" [5] . In an ideal situation, the preservation policies will guide the preservation activities in an organisation. As the field in which the organizations act is rapidly changing, and the insights in digital preservation change, the preservation policy documents should be a regularly revised and updated.

The European project SCAPE has designed a Preservation Policy Model that will support organizations to build their preservation policy documents. Before this, several European projects investigated preservation policies. These results are input for the current work in the SCAPE project.

The DL.org project investigated "interoperability" as an important means to enable digital libraries to get the most value out of their collections and to enable "sharing" and "building by re-use". By being "interoperable" on various aspects, it would be possible to share collections and to collaborate between organisations. Digital libraries is here more broadly defined, not restricted to digital libraries in a traditional sense, but to "a potentially virtual organisation, that comprehensively collects, manages and preserves for the long depth of time rich digital content, and offers to its target user communities specialised functionality on that content, of defined quality and according to comprehensive codified policies [4]. One of the areas for interoperability identified in this report is "preservation policies", for which the DL.org project designed a conceptual approach.

The PLANETS project introduced the "preservation guiding document" [6] including a conceptual model and a vocabulary for preservation guiding documents. The key focus was the digital collection and the risks that might threaten that collection. The

[1] http://www.alliancepermanentaccess.org/index.php/knowledge-base/member-resources/digital-preservation-glossary/

preservation object, within a digital collection, has characteristics and lives in an environment. The identification of a preservation risk will lead to a preservation action, that takes into account the characteristics of the object and the environment in order to formulate requirements.

The Shaman project defined a number of catalogues and processes needed in digital preservation from the business governance viewpoint, such as a Policy Catalogue that provides a list of all the preservation policies, a Driver/policy/goal/objective Catalogue that provides a breakdown of preservation drivers, policies, goals and objectives within the organisation. Further a Contract/measure Catalogue: providing the list of all policies and associated strategies and finally the Preservation Management Processes representing the processes which manage the preservation in the organization [1].

The SCAPE project is dedicated to the challenges of large scale, heterogeneous collections of complex digital objects. The digital objects are held in the collections of various participating content holders, like libraries, web archives and data centres. The scale of these digital collections implies that preservation activities that need to be performed will limit the possibility of manual involvement, and require more automation through the use of workflows and high-performance systems. Preservation activities need to be guided by a preservation policy.

The SCAPE project will run until 2014. The experiment described in this article is an intermediate result that gave us input to shape further work. The scope in this experiment has been limited to preservation policies that are relevant for preservation watch and preservation planning.

2. PRESERVATION AREAS

Preservation Policies will guide Preservation Actions. In digital preservation however, a preservation action will often be preceded by an identified risk, based on monitoring several areas of interest, and a combination of the outcomes leading to a decision to act. The identification of the most appropriate action is done in the Preservation Planning process, which produces a preservation plan. Enacting the preservation plan will result in the Preservation Action. In SCAPE the Preservation Watch area will be enriched by the SCOUT system [9]. SCOUT is an automatic preservation watch system that will detect preservation risks and opportunities. The Preservation Planning will be extended by new versions of Preservation Planning tool PLATO[2]. In both cases, a detailed level of preservation policies will be needed to enable the planning and watch services to act according to a specific set of institutional preservation policies.

2.1 Preservation Watch

In the Planets project an extension of the OAIS model was designed, the Planets Functional View [14], in which special attention was paid to a Preservation Watch function that brings together several monitoring functions. One could imagine that in case of large collections, not all the areas to be monitored can be covered by activities, done manually by humans. Instead an organisation should identify which elements should be monitored and this information could then be fed into an automatic monitoring system. The focus will be determined by the content of the preservation policies. Take for example a preservation policy that would limit the diversity of file formats that an organisation is willing to accept. Monitoring the developments related to file formats can then be restricted to the file formats that are allowed and subsequently be automatically monitored.

2.2 Preservation Planning

Preservation Planning is another area where preservation policies provide important input. If one wants to plan preservation actions that can support the long term preservation of a digital collection, input for this process should come from the preservation policies that are related to the digital material as defined by the organisation and its goals [3].

3. SCAPE PRESERVATION POLICY MODEL

3.1 Policy levels

The SCAPE Preservation Policy Model consists of three preservation policy levels that will support an organisation to create their preservation policies set. By connecting these three levels and identifying clearly which level is fit for which purpose, we intend to make the creation of a preservation policy for organizations more straightforward.

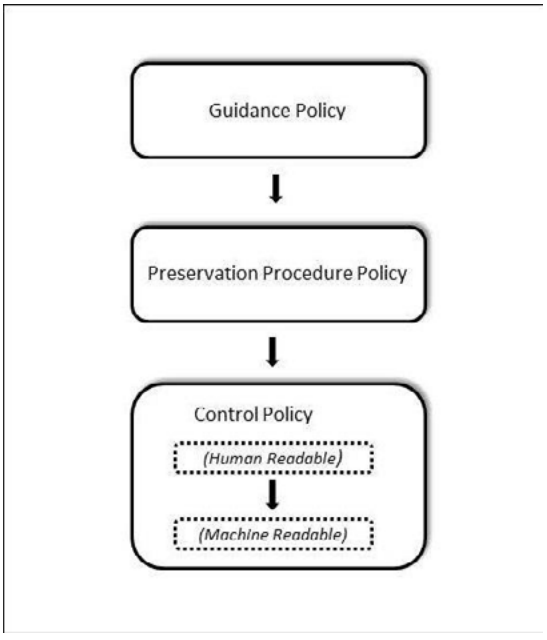


Figure 1 SCAPE Preservation Policy Model

The three levels of policies identified in SCAPE are:

1. **High level or guidance polices.** On this level the organisation describes the general long term preservation goals of the organisation for its digital collection(s). One example is that an organization decides to act according the OAIS model.

[2] http://www.ifs.tuwien.ac.at/dp/plato/intro.html

2. **Preservation Procedure policies.** These policies describe the approach the organisation will take in order to achieve the goals as stated on the higher level. They will be detailed enough to be input for processes and workflow design but can or will be at the same time concerned with the collection in general. These are likely to be made publically available.

3. **Control policies**. On this level the policies formulate the requirements for a specific collection, a specific preservation action, for a specific designated community This level can be human readable, but should also be machine readable and thus can be used in automated planning and watch tools to ensure that preservation actions and workflows chosen meet the specific requirements identified for that digital collection. These are likely to be kept internally within the organisation.

It is the interaction between the Preservation Procedure level and the Control Policy level that is the focal point of study. How much information is enough to transform the decisions and statements in the Guidance Policies and the Preservation Procedure Policies into actionable Control Policies.

3.2 Control Policy Model

The control policies created through the translation of natural language policy are intended to capture the whole policy intent, enabling automatic checking of the state of the world in watch or potential preservation plan in planning. They provide the local organisational environment within generic tools and ensure that these automated tools are not concerning themselves with areas which the organisation is not interested in; honing the tools to the specific circumstance. By using a standard model to represent this information, then two separate tools can use the same policy basis to achieve different aims enabling policy interoperation. This is the SCAPE Control Policy Model (figure 2.)

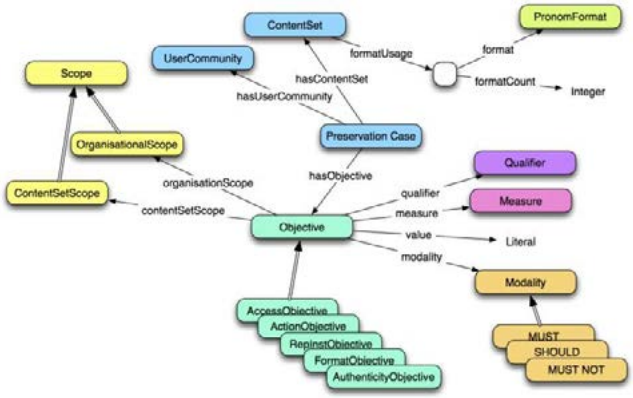


Figure 2 Overview of Control Policy Model

The SCAPE Control Policy Model provides a controlled vocabulary or set of terms and relationships that allow for the description of policies. A key aspect here is that the control policies are expressed in a, unambiguous, machine readable way, rather than as natural language. A policy that states (in English) that "Most formats used must be ISO standardised" is potentially open to interpretation -- what do we mean by "most formats" or even "ISO standardisation"? The controlled policy vocabulary provides a common set of terms that can be used, and on whose interpretation there is a shared agreement. The states of affairs that the objectives define and describe can then be tested or evaluated through some automated processes (without an agreement on the interpretation of terms it is very difficult, if not impossible, to automate this). For example, the policy above states that most formats used for a particular content set must be ISO standardised. A content profiler, such as the c3po tool[3], can analyse document collections and provide information about the formats used in that collection. Format registries (e.g. PRONOM[4]) provide detailed information about the characteristics of formats. By integrating all this information along with an unambiguous interpretation of the policy, the conditions expressed in the policy can be automatically checked, and suitable actions planned. Further advantages of a machine readable policy expression include the ability to validate or check for conflicting or subsuming policies.

The Control Policy Model provides vocabulary that is used to describe particular domain entities: situations, formats, content sets etc. Key entities described in the model are Content Sets, Objectives and Preservation Cases. A Content Set represents a collection of objects that are the focus of the policy. Objectives are the atomic building blocks of the policies. In general, an Objective will refer to a property (see below) along with a value for the property and a Modality that indicates whether or not the expected value is an absolute requirement or prohibition, expressed as MUST/MUST NOT/SHOULD etc.[5] Objectives are generic in that they describe states of affairs without referring to specific content sets or organisations. This facilitates the sharing of Objectives across policies. A Preservation Case ties objectives to a Content Set and intended User Community. Objectives may refer to properties that representations of content have; properties of the formats themselves; tools used and so on.

The properties in Objectives are taken from a collection of measures[6] -- properties that describe particular characteristics of items, formats or actions. For example, "Number of free tools that are open source"[7] is a measure that gives some indicator for the adoption of a format. Measures are further organised into "attributes"[8] -- collection of measures relating to particular characteristics and "categories"[9] -- high level groupings of attributes. A number of measures have been defined by the SCAPE project. In the future we expect measures to be shared across communities -- improving opportunities for sharing and exchange of practice. It may also be the case that particular domains or organisations will want to define their own particular measures -- extending the vocabulary in this way is possible.

Note that the model is simply there to enable the objectives to be stated in an unambiguous way. The model itself does not attempt to check whether or not the statements are true. Such checking will be done by other tools (for example the PLATO planning tool). Further details of the policy models and their use in the SCAPE preservation ecosystem are discussed in [11].

The Control Policy Model of SCAPE uses the W3C's family of representation languages. The models are defined as OWL [12]

[3]http://ifs.tuwien.ac.at/imp/c3po

[4] http://www.nationalarchives.gov.uk/PRONOM/

ontologies, with particular objectives being represented as an RDF [13] knowledge base. This use of standardised representations allows the possibility of existing tools to support the creation, management and manipulation of the policy instances.

Tools that support the user in defining policies using the control policy model are essential -- we cannot expect users to work directly with representations such as RDF. The model itself assists in this process as it can provide constraints as to what users can express, controlling and focusing the expression of the policies. A prototype web application that supports the user in defining objectives has been developed. As we discuss below, however, the process of moving from a high level expression to the specific control policy elements is non-trivial.

4. Verification of the Model using two real life Policies

Having defined the SCAPE policy model, we have verified this approach by using existing policy documents from two of the SCAPE partners to create control policies, both in human and machine readable forms

We used the policies of the State and University Library Denmark and the ISIS Data Management Policy of the Science and Technologies Facilities Council.

Although these policies could not strictly be categorized as either a Guidance Policy or a Preservation Procedure Policy, they were the currently available information with respect to the preservation intentions of both organizations and would reflect the situation in many organizations.

4.1 Policies at the State and University Library

A few years ago the State and University Library created a Digital Preservation Policy (DP Policy[7]) and a Digital Preservation Strategy [8]. The DP policy is at a very high level declaring the purpose and scope of the State and University Library's digital preservation. The DP Policy works at a management level and consists of very general statements. It is revised once a year.

In addition to this policy the State and University Library developed a DP Strategy. This details the high level policies formulated in the DP Policy and is concerned with the overall collection management. It does not specify anything about specific collections but defines how to make the right decisions according to the State and University Library policies. For instance the DP Strategy does not specify precisely what format to use for a specific collection, instead it states that the choice of format for a specific collection must be in line with the policies in the DP Strategy, in the case of formats it must be an open format, it must be well-documented etc.

The DP Strategy is the link between the high level policy, and the preservation plans that have been developed at the State and University Library for specific collections. The collection specific preservation plans transform the policies on the Preservation Procedure Level, in case of the State and University Library the DP Strategy, into human readable control policies that, combined with the general statements from the DP Strategy, form the basis for developing machine readable Control Policies.

In SCAPE The State and University Library has performed an experiment with transforming DP Strategy on the Preservation Procedure Level and the collection specific preservation plans into machine readable Control Policies.

4.2 Policies at the Science and Technology Facilities Council (STFC)

STFC's high level, organizational wide Data Policy [15] states that underlying data should be kept for at least ten years after the end of a project or in perpetuity if it is unrepeatable observational data and that all data should have a Data Management Plan. This data management plan should address preservation as part of the data lifecycle, the focus within STFC is on data management rather than preservation due to the nature of STFC's business which is supporting the processes of creating new scientific data and ensuring this remains useable.

The ISIS Neutron Spallation Source, one of the large scale scientific facilities provided by STFC has a Data policy for users of the facility [10]. Although this is not exclusively concerned with preservation, it addresses some of the topics covered in preservation procedure policy and has been used as the starting point for the creation of control policies to support the Research Data Testbed scenarios provided by STFC elsewhere in the SCAPE project.

4.3 Applying the model to a real life situation

To enable to generation of control policy statements which can be used elsewhere in the SCAPE project a process of elaborating these statements needed to be identified. There are two key differences between policy aimed at a human audience and policy to be used automatically:

There is a difference in intent and viewpoint between written, human readable policies, especially at the higher levels and the control level policy. High level policy is trying to set the boundaries of acceptable states whereas control level policy is aiming to be precise in defining conditions for those states

The second difference is the implicit/explicit dilemma. A person will need less documented facts as they can use other implicit information, whereas a computer system only knows what it is told. Being able to ensure all implicit information is made explicit is a hard task to undertake.

4.3.1 Process for creation of control policies

There are two possible starting positions: (1) that the natural language control level policy is already documented and (2) that natural language preservation procedure level policy exists but natural language control level policy is implicit and is not contained in a single document describing detailed preservation decisions for the collection. For our experiments both of these states applied.

During the experiment we identified the following stage and steps. The three stages are (1) steps which apply to the whole policy

[5] cf RFC 2119 <http://www.ietf.org/rfc/rfc2119.txt

[6] http://purl.org/DP/quality/measures

[7] http://purl.org/DP/quality/measures#139

[8] http://purl.org/DP/quality/attributes

[9] http://purl.org/DP/quality/categories

document, (2) steps which need to be applied to each policy statement and (3) final review of the results.

Whole Policy Steps

1. **Define the content set that the policy addresses**

The content set is an intellectual cohesive collection of digital objects to which all the objectives within a preservation case apply.

The differences between the two organisations showed clearly a different approach in identifying the collections, for STFC the policy created a single content set related to the way the data were created and collected, and at SB the collection was a heterogeneous set of Radio Television Collection, as the policies were written on this level and reflect the organisation's view of their information. It should be noted that the STFC ISIS formats are specialised and consist of a local format for early data and a domain specific format for later data, and so for data management purposes there is no need to further divide the data; however for preservation purposes where we are interested in the semantics within the files, then there may be a need to describe collections in a different manner.

2. **Identify the user community/ roles required by the policy**

It is important to be able to identify who will be enacting the policy statement. Although the SB and STFC user communities identified had different names, they both were aligned to the DL.org [1, p.23] End Users which identifies three types: creators, consumers and administrators.

3. **Map policy statements to high level concepts**

To assist in identifying the risk or preservation case that the particular policy statement addresses, it is mapped to one (or more) of the high level concepts we already identified in SCAPE.

So the ISIS Data Management policy fragment **"3.1.1** ***All raw data will be curated in well-defined formats for which the means of reading the data will be made available by the Facility",*** maps to the high level concepts of format and access and so the final preservation case will be concerned with these aspects.

Steps for each line of policy

1. **Clarification to implicit meaning**

This stage is designed to ensure that the natural language version being worked on does not have any "hidden" meaning within the words.

2. **Identification of Control Policy Model Preservation Case**

A Preservation Case ties Objectives to a Content Set (defined in step 1) and intended User Community (defined in step 2) This step should assist in identifying a particular Preservation Case for this particular policy statement.

3. **Identification of Objectives for this content set**

The Objectives are the measurable machine readable statements to be generated from the policy fragment being considered. These for example can be access objectives *rendering tools should exist for specific environments in use by the user community* or file format objectives *only ISO standard file formats should be in the collection*. The Objectives need to be phased in clear statements (MUST, SHOULD, >, < etc.)

4. **Generate control statements**

Tooling with a GUI will support the end user to create the machine readable control statements; in this case we use a set of already created attributes and measures.

Review the Preservation Cases

1. **Review the preservation cases identified**

Having completed the whole policy, then a check should be made as to whether any control policies and/or preservation cases overlap and whether it might be advisable to merge the outcomes or identify those which apply to the whole organisation.

5. CONCLUSIONS

Several conclusions can be drawn from this experiment. Firstly, it is possible to create machine readable control policies based on existing policy documents and using the Control Policy Model described in this document. The ease of doing so depended on the level of policy documents and the familiarity of the creator with the preservation intent and specific collection knowledge and in both cases the policy documents were too generic and detailed information needed to be gathered from other sources. This process also assumes that all relevant topics will be covered in the Preservation policies; there may be occasions where the control policy may come from another source – such as a specific requirement of the software used.

There are two main challenges still be to be worked on. The first is that the process moving from the often implicit to the explicit; is in practice a difficult task and the requirement to make control policies unambiguous may not be achievable for all policy elements. Secondly the granularity of the preservation case is still under discussion. The preservation case groups the objectives, content set and users together around the mitigation of a risk and will be used in the Watch and Planning tools. What is the appropriate level of granularity working from the policy, may not be the same as that required for Watch or for Planning for a Preservation Action. Both of these use triggers to action and the linkage between these and preservation cases are still under discussion. Currently we suggest that as it is not easy to identify the right level of granularity when defining control policies, we recommend to creating fine distinctions first and merging categories during the final stage.

This process leads from the natural language to machine readable policy, there is no process available to check that this machine readable policy is actually the same intent as the natural language policy, although ensuring specific linkages/relationships to be made between statements in the two levels would assist in this. Further development of a catalogue of policy elements related to the controlled vocabulary will contribute to solving these problems.

6. ACKNOWLEDGMENTS

This work was partially supported by the SCAPE Project. The SCAPE project is co-funded by the European Union under FP7 ICT-2009.4.1 (Grant Agreement number 270137).

7. REFERENCES

[1] Antunes, G, Barateiro, J, Becker, C et all: *Shaman Reference Architecture. Final version – update year 4* [2012. Retrieved 22-04-2013 from http://shaman-ip.eu/sites/default/files/SHAMAN-REFERENCE%20ARCHITECTURE-Final%20Version_0.pdf

[2] Bradner, S: *Key words for use in RFCs to Indicate Requirement Levels*. RFC 2119. 1997. Retrieved 22-04-2013 from : http://www.ietf.org/rfc/rfc2119.txt

[3] Becker, C., Kulovits, H., Guttenbrunner, M., Strodl, S., Rauber, A., & Hofman, H. (2009). Systematic planning for digital preservation: evaluating potential strategies and building preservation plans. *International journal on digital libraries*, *10*(4), 133-157.

[4] Candela, L and Nardi, A (ed.) *The Digital Library Reference Model* , 2011 p. 17. Retrieved 22/04/2013 from: from http://www.dlorg.eu/index.php/publications ver, Crafty Content

[5] Candela, L. and A.Nardi (ed.) *Digital Library Technology and Methodology Cookbook* p. 68 Retrieved 24-04-2013 from http://www.dlorg.eu/index.php/publications

[6] Dappert, A: *Report on the Conceptual Aspects of Preservation, Based on Policy and Strategy Models for Libraries, Archives and Data Centers*. Planets Project, 2009. Retrieved 22-04-2013 from http://www.planets-project.eu/docs/reports/Planets_PP2_D3_ReportOnPolicyAndStrategyModelsM36_Ext.pdf

[7] *Digital Preservation Policy for the State and University Library Denmark*. 2012 version 2.0 Retrieved 22-04-2013 from: http://en.statsbiblioteket.dk/about-the-library/ddpolicy

[8] *Digital Preservation Strategy for the State and University Library*, Denmark Version 2.0 June 2012 Retrieved 22-04-2013 from:http://en.statsbiblioteket.dk/about-the-library/dpstrategi

[9] Faria, L., P. Petrov, K. Duretec, C. Becker,M. Ferreira, and J. C. Ramalho. Design an architecture of a novel preservation watch system. In: *International Conference on Asia-Pacific Digital Libraries (ICADL)*. Springer, 2012

[10] *ISIS Data Management Policy*. Retrieved 22-04-2013 from http://www.isis.stfc.ac.uk/user-office/data-policy11204.html

[11] Kulovits, Hannes, Kraxner,Michael, Plangg, Markus, Becker, Christoph, Bechhofer, Sean. *Open Preservation Data: Controlled vocabularies and ontologies for preservation ecosystems* 10th International Conference on Preservation of Digital Objects (iPRES 2013), 2013

[12] *OWL 2 Web Ontology Language Document Overview* (Second Edition) 2012. Retrieved 22-04-2013 from: http://www.w3.org/TR/owl2-overview/

[13] *Resource Description Framework (RDF*) Retrieved 22-04-2013 from: http://www.w3.org/RDF/

[14] Sierman, B. and Wheatly, P: *Evaluation of Preservation Planning within OAIS, based on the Planets Functional Model. Planets Project* 2010 Retrieved 22-4-2013 from http://www.planets-project.eu/docs/reports/Planets_PP7-D6_EvaluationOfPPWithinOAIS.pdf

[15] *STFC scientific data policy* Retrieved 22-04-2013 from http://www.stfc.ac.uk/Resources/pdf/STFC_Scientific_Data_Policy.pdf

Analysis of the variability in digitised images compared to the distortion introduced by compression

Sean Martin
British Library
Boston Spa
Wetherby, LS23 7BQ, UK
+44 1937 546716
Sean.Martin@bl.uk

Prof Malcolm Macleod
QinetiQ Ltd
St Andrews Road
Malvern, WR14 3PS, UK
+44 1684 543796
mdmacleod@iee.org

ABSTRACT

The paper evaluates the noise which is present in digitised images of very high quality and the noise/error which results when such images are compressed and then decompressed. The variations between pairs of captured images of identical material were compared and the two best pairs of images were identified. The variations between these pairs were then compared with the variations introduced by compression and decompression of those images. We found that even lossy compression can result in significantly lower variation than that between the best pairs of original images caused by imaging noise. We report the results of a qualitative questionnaire which are in good agreement with the quantitative assessment. The conclusions suggest that given the extent of noise in the imaging process the current practice of storing lossless master digitised images could be replaced by the use of more compact compressed images, arguably with no loss of quality.

Keywords

digitisation, camera, scanner, digitised image, camera noise, JPEG 2000, image compression, PSNR

1. INTRODUCTION

The motivation for the work described in this paper arose when some simple experiments were conducted in a digitisation studio. A particular item was imaged several times and the resulting images were examined visually. The extent of the differences between the images at a detailed level was surprising, and this prompted further investigation. This evolved into the structured process which is described in this paper. Meanwhile a review of the literature, summarized in Section 2, identified several papers which discuss noise in the imaging process and its consequences.

Three physical items were each imaged in colour with seven devices that produce images that can be compared automatically. Each item was imaged five times in rapid succession with the same device, a camera or a scanner, without moving the item, and without changing the background lighting. The variations between 210 pairs of these images were assessed. Additionally, two selected "best" pairs of images were subject to detailed further examination. One pair was produced by a top of the range camera, and the other by a regular production camera.

The differences between these pairs of images were compared with the variations caused by compressing one of the images in each pair in a lossy manner. It was found that for modest amounts of compression, the variations introduced by compression were less than the variations between the original lossless master images.

This quantitative assessment was complemented by a qualitative questionnaire in which images were compared by eye and respondents were invited to indicate which pairs of images had least or most differences. The qualitative assessment produced results in line with the quantitative assessment. The questionnaire also included examples where greater compression was applied and respondents were asked to indicate whether the resulting images were considered perfect, acceptable, marginal or unacceptable. It was found that modest compression could be applied without compromising the perceived visual quality of an image, even when the image has been greatly magnified.

This final part of the questionnaire produced an interesting result. In some cases the alternative lossless original master images were deemed merely acceptable, whereas compressed images with very small loss were deemed to be perfect.

These results, particularly the last one, question the need for retaining digitised images in a lossless manner. It is clear there are noise-induced differences between original high quality master images, while a mild level of compression can result in much less variation. When applied in an appropriate manner this could reduce storage costs, conservatively by 30-70% compared with storing lossless JPEG 2000 files. For bulk digitisation greater cost saving is possible. The choice might be dependent on the subject matter but a strong case is put forward that a minimum of the order of 30% compression is achievable with little reduction in perceived quality or value.

2. NOISE AND IMAGING

2.1 Noise in the Imaging Process

While noise in imaging is discussed widely in the literature, there has been limited attention regarding the extent and nature of noise that occurs in the imaging process and is thus present in digitised images.

(Liu, et al. 2008) [9] states that there are five primary noise sources in a camera with a CCD (charge coupled device) sensor. These are: fixed pattern noise (FPN), dark current noise, shot noise, amplifier noise and quantization noise. These arise in the successive processes by which photons cause electron activity, which is amplified and then digitised - noise is introduced at each stage. The paper discusses the statistics of noise and how noise can arise in the colour that is recorded - known as colour noise. (Faraji and MacLean 2006) [5] describe signal-independent noise and signal-dependent noise, and they characterise noise sources in a similar way to [9] including photon noise, FPN, amplifier noise and readout noise. They refer to an extensive discussion of noise in (Janesick 2001) [6] and they also note that at low light levels the noise is independent of the signal, at mid light levels the noise becomes signal dependent – arising from shot noise, photon noise and dark noise, typically with Poisson distributions. At high light levels FPN proportional to the signal dominates. (Chen, et al. 2009) [2] also characterise noise as FPN and random noise.

(Kurosawa, Kuroki and Akiba 2009) [8] establish that it is possible to identify that an image, or more specifically a series of images in a video, were taken by a particular camera. Distinctive FPN can be produced by individual "hot" pixels and the spatial arrangement of these pixels can be recognised in an image, and the camera thereby identified. The ability to identify the camera from its noise signature is analogous to identifying a gun from a bullet fired from it.

(McHugh) [10] gives an excellent tutorial on noise in digital cameras, and states that digital cameras produce three types of noise: random noise, FPN, and banding noise, noting that the latter is highly camera-dependent. The following example pictures from [6] are reproduced by permission:

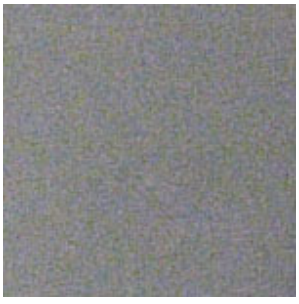


Figure 1: Example of random noise

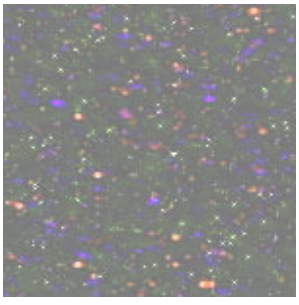


Figure 2: Example of fixed pattern noise

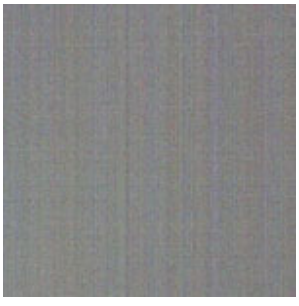


Figure 3: Example of banding noise

[10] also observes that noise is more prominent in darker regions, and that noise can comprise fluctuations in both colour and luminance, where, for example, chroma noise can be evident as colour superimposed on a grey portion of an image. Noise can be both fine- and coarse-grained in texture.

The signal to noise ratio (SNR) is a useful and universally used way of comparing the relative amounts of signal and noise in any electronic system; high ratios will have very little discernible noise whereas the opposite is true for low ratios.

The literature concerning noise and images arises from a wide range of disciplines, including astronomy with low light levels, and medical imaging, such as (Belbachir and Goebel 2006) [1] which discusses noise in the incoming photon stream. Many of the cited papers discuss schemes for reducing noise, and they therefore discuss the sources of noise and models for it.

The concept of noise in a camera image is for some an abstract notion. However, it may be helpful to relate it to the hiss heard in an old audio recording. The hiss is noise – if the record is replayed then the hiss could be different, although the symphony may sound the same. The value is in the symphony, whereas it is rarely of any value to record and reproduce faithfully the hiss that occurred on one particular occasion.

2.2 Image compression and noise

Image compression is an important technology for reducing the amount of storage required to hold images, or the communication capacity required to transmit them.

There is a widespread opinion in the library and archive community that it is vital that images be stored losslessly. The noise in such an image would also be preserved. However, as is well known, random noise is inherently difficult to compress. This can lead to a significant proportion of a lossless compressed digitised image file being used to reproduce exactly the noise in the image.

If instead lossy compression is used, then the decompressed image will differ from the original. Since the compression is lossy it is likely that it will fail to encode the noise completely, as it is difficult to compress. However, with a low degree of compression, and hence loss, it is likely that the signal in the image will remain almost intact other than a small amount of distortion that is introduced. Provided the power (or extent) of the distortion that is introduced is less than the power of the original input noise, it can be argued that the decompressed image has exactly as much quality (SNR) as the original. This hypothetical consideration is not exactly what occurs, but it demonstrates the argument that the artefacts in an image reconstructed after compression may represent no loss of quality compared to an original master image with the unavoidable noise present in it.

To explore this hypothesis requires a detailed comparative analysis of input noise and the noise resulting from compression. It is that which is the goal of this paper.

3. EXPERIMENTAL WORK

3.1 Introduction

Initial experiments were conducted in which the same item was imaged several times in a manner designed to be as close to identical as possible. The resulting images were compared and there were visually obvious significant differences between the magnified images. These experiments led to the development of a systematic process for characterising the noise in an image, as described below.

Each original image created in this process is a lossless master image file and therefore 'authentic', but the images are different from each other because of the presence of noise. We developed a method whereby we could compare (a) the variations between these lossless master files with (b) the variations, usually called degradation, introduced by compressing a master file in a lossy manner.

The full detail of the process is now described.

3.2 Method of Imaging

Three separate physical samples were selected and the same samples were imaged multiple times with different cameras and scanners that are, or were, in regular use in the digitisation studios in two national libraries, the Norwegian National Library and the British Library. The three samples that were imaged are approximately A3 in size.

A sample was placed under a camera or scanner and imaged multiple times in quick succession with no deliberate change in the background lighting conditions in the digitisation studio.

Each sample was imaged in this way with ten different cameras or scanners. However, the images from three devices were later discarded since their images were too variable to make detailed comparisons practicable. This left sets of high quality images from seven devices which were appropriate for detailed inspection. These included one scanner and six cameras, all in regular production use. Two were automated page turning machines, and each of those had two standard professional Phase One backed digital cameras. Images taken with these devices were designated as N01-N04. A top of the range Hasselblad specialist camera was designated as N05. The scanner was designated as N06, and a separate Phase One backed digital camera was designated as B07. An N indicates that the images were taken at the Norwegian National Library, and B indicates that the images were taken at the British Library. As will be seen later there is broad consistency of quality between the best pairs of images produced by these devices.

However, these sets of images also showed detectable variations and so we undertook more experimentation on the manner in which the images were taken. The nature of the variations is discussed later.

We experimented by taking multiple images with a longer (ten second) delay between them – to see if the action of imaging introduced a small vibration that caused a wobble in the image. However, the resulting variations were similar to those in an original set of five images.

We conjectured that there might be a lensing effect arising from density variations in the air flow between the camera and the item. We therefore set up an experiment where one half of the item had air blown across it with a fan while the other half had no air flow. We found that the two halves of the image had similar variations to an original set of five images, and we could detect no differences caused by the difference in air flow.

We emphasise again that each image we used in the experiments described below is an example of what an archivist would regard as a 'valid master file'.

3.3 Quantitative Assessment of Images

Within a set of five images of the same item taken with the same camera or scanner, ten pairwise comparisons are possible. (The first image is compared with four others, the second image with three others, and so on.)

As explained above, three separate items were imaged, and seven devices used to produce sets of images for automated quantitative assessment.

There were therefore 210 pairwise comparisons of images available. An immediate impression was that for each pair of images there are significant variations between them, despite each image being an authentic master image.

The experiments assessed the similarity between a pair of images. However visual inspection showed that there were often small, but quite noticeable, lateral shifts between images, and this greatly complicated the comparison process. The comparison thus had to be preceded by aligning the two images to obtain the best correlation score.

A simple hill climbing technique proved effective for correcting shifts that were small compared with the size of features in the image. This worked well on the images from the selected seven devices. Reference was made earlier to devices that delivered images sufficiently different to make comparison difficult – in one case because the observed shifts were large, for example 70-100 pixels, whereas a typical feature might only be ten or so pixels across. A simple hill climbing algorithm was no longer effective since it stopped at intermediate local maxima and failed to find the overall best fit. In another case the device produced images whose width and height dimensions were so significantly different as to make comparison difficult.

A simple PSNR (peak signal to noise ratio) was used as the correlation metric, though other metrics are possible and have been reported to produce better comparisons between digitised images. As is customary, the PSNR is expressed in logarithmic (deciBel, dB) units, which give the best correspondence with the perceived quality.

It was often also found that the lateral shift was not constant and could vary by a small amount across the image – this is a form of spatial distortion between a pair of images. Often there is a slow progression, with the lateral shift slowly changing or drifting across the compared images. We also observed one case where the extremities of the compared images diverged – in effect there had been a small change in the magnification.

The method of comparison took the lateral shift into account by considering a portion, or tile, from each image in turn and then aligning and correlating each pair of tiles independently. A PSNR metric was then calculated for the entire image as an aggregate of the metric from each optimally aligned pair of tiles. A typical tile size used was 400 x 400 pixels. A tile size of 100 x 100 was also tested and produced similar results.

The lateral shifts were usually not an exact whole number of pixels. The alignment technique was therefore extended such that once there was optimum alignment based on shifts of a whole number of pixels between a pair of tiles, each tile was then expanded by interpolation, and then aligned to an accuracy equivalent to a fraction of a pixel in the original image. Early experimentation showed that a bi-linear interpolation was as effective as bi-cubic interpolation, and so bi-linear interpolation was used for this further analysis.

With this enhancement, each of the 210 pairs of images was aligned to 0.25 pixels.

3.4 Quantitative Assessment Results

Table 1 shows a summary of the comparison of pairs of images. The first column cites the identity of the imaging device, referred to as N01-N06 or B07. The six remaining columns record for each of the three sample documents A, B and C the PSNR results. “Av” is computed by averaging the PSNR values from the ten pairwise comparisons (not in dB form) and then converting the average to dB. “Max” is the maximum within the set. A standard colour coding has been applied to help highlight particular scores where red indicates a low score and blue a high score.

Table 1: PSNR in dB comparing images without shifting

Sample	A	A	B	B	C	C
	Av	Max	Av	Max	Av	Max
Device						
N01	30.790	36.742	31.400	37.024	36.742	36.942
N02	32.712	36.431	36.874	36.925	36.390	36.656
N03	30.083	37.042	32.009	38.277	32.048	37.916
N04	30.095	37.373	30.391	37.823	36.123	37.348
N05	41.449	42.128	42.317	43.000	42.197	42.508
N06	29.851	31.286	28.011	31.335	29.669	30.961
B07	19.479	33.842	22.878	38.862	16.687	36.347

We see that:

Device N05 has consistent and relatively high scores. For each of the three samples the maximum for N05 is only a little greater than the average – this indicates that the ten pairwise comparisons are quite consistent. The N05 scores are also consistent across the three samples A. B and C.

By contrast device B07 shows much greater difference between the average and maximum scores; for example an average of 16.687dB and a maximum of 36.347dB for sample C. This indicates considerable variation between the individual scores, as will be confirmed later.

Devices N01 to N04 are all supplied by the same manufacturer. The consistency of their images falls between those for B07 and N05, with a greatest difference between average and maximum of around 6dB and the least being only 0.05dB.

Table 2 shows the results of comparing matching tiles from pairs of images. The tiles were processed independently within each pair of images in the manner previously described, where the tiles from the different images were aligned for best fit to the nearest pixel and an aggregate PSNR value was derived for the entire image.

Table 2: PSNR in dB comparing images after shifting

Sample	A	A	B	B	C	C
	Av	Max	Av	Max	Av	Max
Device						
N01	30.811	36.742	31.400	37.024	36.742	36.942
N02	32.712	36.431	36.874	36.925	36.390	36.656
N03	31.020	37.042	33.780	38.277	33.395	37.916
N04	32.424	37.373	30.494	37.823	36.123	37.348
N05	41.449	42.128	42.317	43.000	42.197	42.508
N06	29.857	31.286	29.478	31.335	29.850	30.961
B07	25.861	33.842	29.048	38.862	24.171	36.347

We see that:

Devices N02 and N05 have identical results in tables 1 and 2 indicating that all their pairs of images are already aligned - there are no lateral shifts between them.

For device B07 the average scores increase from table 1 to table 2 – this demonstrates that the shifting algorithm is able to improve the alignment between some pairs of images. However, for B07 and the other four devices the maximum scores remain unchanged indicating that the pairs of images which generated them were already optimally aligned.

The information from table 2 is summarized in table 3 which records three scores for each of the devices. These are the average of the averages for the three items A, B and C, the average maximum for the three items, and finally the overall maximum value.

Table 3: Average and Maxima from Table 2

	Average of Averages	Average Maximum	Maximum of Maxima
Device			
N01	32.984	36.903	37.024
N02	35.325	36.671	36.925
N03	32.732	37.745	38.277
N04	33.014	37.515	37.823
N05	41.988	42.545	43.000
N06	29.728	31.194	31.335
B07	26.360	36.350	38.862

The last column shows quite consistent results. Devices N01 to N04 are similar, with maximum scores of 36.9 to 38.3dB. Device B07, which as noted earlier showed considerable variations, had a slightly better maximum score of 38.9dB. Device N05 shows the best results at 43.0dB. All these devices were cameras, whereas device N06 was a scanner. It has a noticeably lower score of 31.3dB.

As reported earlier, visual inspection of pairs of images showed that there was still a discernible shift between pairs of images even though the images were aligned to the nearest pixel. As explained we therefore interpolated pixels and repeated the alignment of tiles within an image, to the nearest interpolated pixel. Table 4 presents a summary of the results when interpolating and shifting were applied to each pair of images.

The results in table 4 show small improvements when compared with the results in table 2. Identical results would not be expected since the basis of comparison has changed. The images for device N05 show increases of around 1.5dB between the two tables for both the average and the maximum scores. The corresponding scores for other devices show greater increases in the range 2-6dB.

Table 4:PSNR in dB comparing images with shifting and interpolation to 0.25 pixel

Sample	A	A	B	B	C	C
	Av	Max	Av	Max	Av	Max
Device						
N01	35.485	38.234	35.804	38.500	38.182	38.418
N02	36.128	37.898	38.354	38.413	37.818	38.141
N03	35.671	38.437	37.115	39.854	37.265	39.478
N04	36.014	38.924	36.040	39.371	37.502	38.894
N05	42.944	43.724	43.757	44.536	43.698	44.070

N06	34.445	34.941	34.670	34.909	34.260	34.570
B07	30.718	35.947	32.183	40.211	27.075	37.162

The best overall individual match was obtained with device N05 and item B where the ten individual comparisons without interpolation between pairs of images are shown in Table 5. These images are designated N05B1-N05B5.

Table 5: PSNR in dB comparing pairs of images using device N05 and Item B

Images	N05B2	N05B3	N05B4	N05B5
N05B1	42.774	42.432	41.711	41.228
N05B2		42.890	42.192	41.724
N05B3			42.747	42.470
N05B4				**43.000**

Table 5 shows that the overall best match pair was between N05B4 and N05B5. This pair was used in later qualitative assessments. As noted earlier these pairs of images are already aligned and so shifting produces identical results. Table 6 shows the corresponding results after interpolation.

Table 6: PSNR in dB comparing pairs of images with interpolation for device N05 and Item B

Images	N05B2	N05B3	N05B4	N05B5
N05B1	44.260	43.865	43.074	42.540
N05B2		44.408	43.617	43.098
N05B3			44.243	43.931
N05B4				44.536

The best match for a standard Phase One backed digital camera was obtained with device B07 and item B. Table 7 shows the ten individual comparisons between pairs of images. The five images are designated as B07B1-B07B5. It is worth noting that some of the other image pairs show significant differences, such as the pair B07B1 and B07B2 which has a remarkably low PSNR of 14.2dB.

Table 7: PSNR in dB comparing pairs of images for device B07 and Item B

Images	B07B2	B07B3	B07B4	B07B5
B07B1	14.168	16.265	16.301	16.130
B07B2		22.401	22.037	22.683
B07B3			30.500	**38.862**
B07B4				29.433

Table 8 shows the results after shifting by integer pixels (i.e. without interpolation). As noted earlier the best match in this set is between B07B3 and B07B5 and its score is not improved by shifting. The score for the poorest image pair (B07B1 and B07B2) has improved but is still significantly below the best value. Table 9 shows the results after shifting and interpolation.

Table 8: PSNR in dB comparing pairs of images with shifting for device B07 and Item B

Images	B07B2	B07B3	B07B4	B07B5
B07B1	25.834	27.667	28.625	27.317
B07B2		28.493	25.330	28.417
B07B3			30.500	**38.862**
B07B4				29.433

Table 9: PSNR in dB comparing pairs of images with shifting and interpolation for device B07 and Item B

Images	B07B2	B07B3	B07B4	B07B5
B07B1	29.529	32.200	33.629	31.097
B07B2		33.230	27.409	34.101
B07B3			30.775	40.211
B07B4				29.645

4. ASSESSMENT OF COMPRESSION

4.1 Analytical work

The previous section identified two sets of 'most similar' images; they were of item B, from devices N05 and B07. These sets were N05B and B07B, and in each set there are five images. Each image was next encoded into a set of JPEG 2000 files with various degrees of compression.

JPEG2000 is becoming increasingly used within the archival community. It supports lossless (reversible) compression using an integer based encoding and also lossy (irreversible) compression using floating-point encoding. The latter can be configured to minimise the loss – where perfect computation would incur no loss but floating point calculations are subject to round off error and this does cause loss. This technique is colloquially known as “minimally lossless”. If lossless integer encoding is taken as a baseline, then minimally lossless encoding typically introduces variations at around 50dB PSNR but with a reduction in file size of 30-40% compared with a lossless JPEG 2000 encoding.

Each image in both sets was encoded in a range of ways: lossless, minimally lossless, and then with a series of lossy compression factors designated as G2 to G12, indicating progressively increasing compression. Each compressed image was compared with the original using PSNR and a compression ratio was derived from the size of the image files. The baseline chosen for the compression ratio was the size of a lossless JPEG 2000 file. There was a particular reason for this. An organisation wishing to store lossless files could choose to use the TIFF format; however, JPEG 2000 offers a lossless format. Those experiencing cost pressure are likely to choose the latter and hence this is an appropriate baseline for determining the additional cost saving in adopting lossy compression. (It should be noted that there are concerns about the ability of JPEG 2000 to retain colour space information; however, when the effect of noise is taken into account it could be argued that a camera is not able to produce a sufficiently accurate colour to make this relevant.)

The compression ratio of a lossless JPEG 2000 file is thus deemed to be 1.0. (A JPEG 2000 lossless file is typically 30-40% smaller than an uncompressed TIFF file.)

Kakadu software was used to encode the images using the British Library JPEG 2000 encoding profile. However the tool used to

derive PSNR and compression ratio used the Leadtools software library to decode the JPEG 2000 images.

Table 10 shows the average PSNR and average compression ratio for each way of encoding the images in each of the two sets N05B and B07B. When two images are identical then the PSNR between them is defined by the PSNR algorithm as infinity; that shows that compression was lossless.

Table 10: Compression ratio and PSNR for N05B & B07B

	Image Set N05B		Image Set B07B	
Compression designation	Compres-sion ratio	PSNR dB	Compres-sion ratio	PSNR dB
lossless	1.00	Infinity	1.00	Infinity
minloss	1.70	50.477	1.59	49.906
G2	1.68	50.255	1.57	49.709
G3	2.24	46.224	2.60	43.745
G4	2.64	44.476	3.06	42.153
G5	3.20	42.836	3.71	40.044
G6	4.26	41.341	4.94	37.685
G7	5.59	39.576	6.48	36.220
G8	7.46	37.231	8.64	34.299
G9	9.94	35.135	11.51	31.952
G10	14.90	32.412	17.26	29.417
G11	19.83	31.566	22.97	28.535
G12	29.73	29.798	34.44	26.738

Table 3 summarised the average and maximum scores for all the devices. The PSNR of the best overall match between a pair of images was recorded there for device N05 as 43.00dB. This lies between the two highlighted rows for device N05 in Table 10.

The PSNR of the best overall match for a standard Phase One backed digital camera was recorded for device B07 as 38.86dB. This lies between the two highlighted rows for device B07 in Table 10.

For device N05 this indicates that a compression ratio of 2.64 produces less variation from an original image than was measured as the best match between a pair of master images as a result of image capture noise. Similarly for device B07 a compression ratio of 3.71 produces less variation than has been measured as the best match between master images.

Visual inspection of the images also confirms that encoded images with less compression than the highlighted amounts have noticeably less variation than the best matching original master files. This forms the subject of the qualitative investigation which is described later.

The minimally lossless images have PSNR values around 50dB. For N05 this is 7.47dB better than the best matched pair of original images, and for B07 this is 11.05dB better. As PSNR is a logarithmic measure this means that the variations introduced by minimally lossless compression are small compared with the variations between these best pairs of original images.

For N05 the root mean square (RMS) variations introduced by minimally lossless compression are 42% of the variations between the most similar original images, and for B07 only 28%.

The information in table 10 is shown in Figure 4 where the two lines characterize the PSNR with increasing compression for the images N05B4 and B07B3. The images for device N05 show a shallower decline than for device B07.

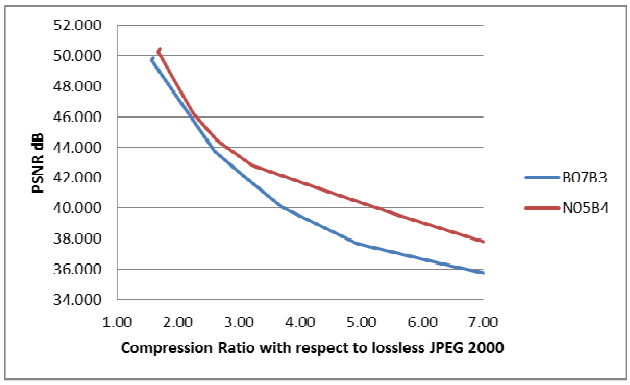


Figure 4: Compression Ratio and PSNR for N05B4 & B07B3 at low compression ratios

The information recorded in table 10 is derived by comparing a compressed file with the original from which it was derived. However, in the set there are a total of five images, and hence there are four alternative master files with which a compressed file can be compared.

So compressed versions of an original image, N05B4, were compared with that original but also with the four alternative master images, N05B1-3, and N05B5. The comparison is in terms of differences as measured by PSNR in dB and the compression ratio with respect to a lossless JPEG 2000 file. The results are shown in table 11.

Table 11: Comparison of compressed versions of image N05B4 with alternative master images for device N05 and Item B

Compression designation	Compression ratio	N05B4 original	N05B1 master	N05B2 master	N05B3 master	N05B5 master
lossless	1	infinity	41.81	42.29	42.83	43.08
minloss	1.70	50.89	41.48	41.98	42.49	42.69
G2	1.68	50.70	41.45	41.96	42.47	42.67
G3	2.24	46.42	40.87	41.31	41.74	41.91
G4	2.64	44.60	40.43	40.83	41.21	41.36
G5	3.20	42.92	39.86	40.20	40.53	40.66
G6	4.26	41.40	39.16	39.45	39.73	39.84
G7	5.59	39.62	38.12	38.35	38.57	38.66
G8	7.46	37.26	36.42	36.57	36.71	36.78
G9	9.94	35.15	34.69	34.79	34.87	34.92
G10	14.90	32.42	32.23	32.28	32.33	32.35
G11	19.83	31.57	31.43	31.47	31.51	31.53

G12	29.73	29.80	29.72	29.75	29.77	29.78

The information from table 11 is also shown in Figure 5. It can be seen that PSNRs of comparisons of compressed versions with the corresponding original start above 50dB and drop fairly rapidly with increasing compression. The best match between master images is between N05B4 and N05B5 at 43dB. Compression by up to a factor of three results in better PSNR than that.

It can be seen from Figure 5 that the PSNRs of comparisons of compressed versions with different master files also drop off with increasing compression, but much more slowly. It had been anticipated there might have been a plateau up to compression by a factor of 3 before this drop off, but it is evident there is an immediate drop off. This can also be seen in comparing the two rows for lossless and minimally lossless compression in table 11. These show a small reduction despite the small changes incurred in using minimally lossless compression.

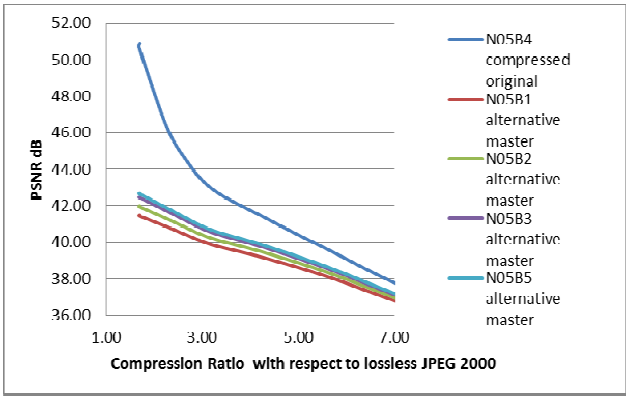


Figure 5: Comparison of compressed versions of image N05B4 with alternative master images for device N05 and Item B

The results in tables 5 and 11 were produced by different tools and there are some minor differences in the results which can be attributed to round-off differences when decoding the images. Kakadu and ImageMagick were used for table 11, whereas Leadtools was used for table 5.

As noted earlier, camera N02 produced consistent sets of five images. The same process was repeated using the five images taken with device N02 and item B. The results are shown in Figure 6.

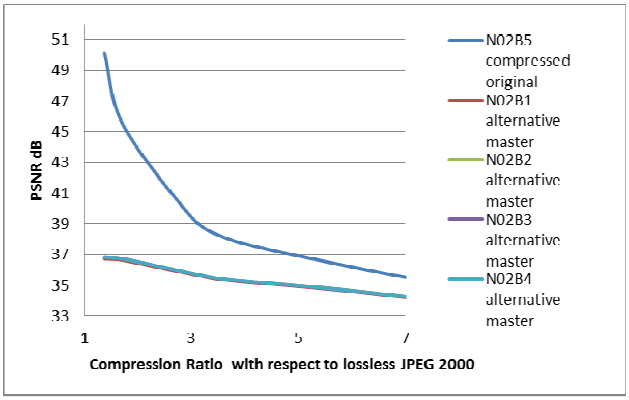


Figure 6: Comparison of compressed versions of image N02B5 with alternative master images for device N02 and Item B

The general patterns of figures 5 and 6 are clearly similar where the four lines for the alternative master files have lower PNSR values, but they are much closer in figure 5.

The comparisons in Figure 6 with individual master images were averaged and these are shown in Figure 7. Also shown are the average PSNR values when compressed versions of image N02B5 are compared with the corresponding compressed versions of the other master files. For example, the G3 compressed version of NO2B5 is compared with each of the G3 compressed versions of the other master files, and their PSNR values are then averaged.

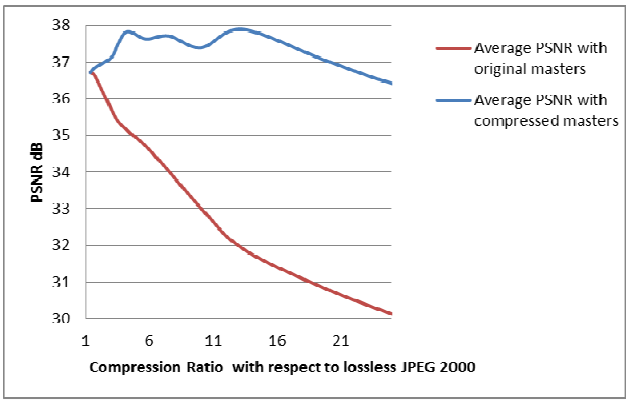


Figure 7: Comparison of compressed versions of image N02B5 with compressed alternative master images and master for device N02 and Item B

Figure 7 shows a remarkable result that the PSNR values rise indicating that the compressed versions are "less different" than the master images from which they were derived. This supports the hypothesis in section 2.2 that predominantly noise is being removed with low levels of compression. With a level of compression above 14 then the signal (or quality) of the image is also being removed. This effect is worthy of further investigation.

4.2 JPEG 2000 Encoding Artefacts

There are several publications which discuss three types of artefacts that can arise when encoding an image using JPEG 2000 at low bit rates and hence with a modest or high degree of compression. These are ringing, colour bleed and tiling artefacts. Ringing and colour bleed can both arise when there is a rapid change such as with a sharp edge or a colour boundary in an image.

(Fang and Sun) [3][4] discuss a technique for reducing the extent of ringing effects that arise from the wavelet compression in JPEG 2000 when encoding at low bit rates. These are visible spurious oscillations or ringing artefacts such as shadows that can occur when there are sharp edges in an image. They show how the technique can be applied and provide examples with levels of compression that result in PSNR values in the range 21-33 dB.

(Nasonov) [11] discusses a method for estimating the extent of ringing in an image and they note that it is as a result of a cut-off of high-frequency information in the encoded image.

(Punchihewa) [12] discusses a technique to evaluate colour bleeding artefacts which result when there is a leakage of colour across distinct colour boundaries in an image.

(Hashimoto et al) [7] discusses techniques for reducing tiling artefacts that can arise in the JPEG 2000 encoding process. Tiles of the image are analysed separately and artefacts can occur at tile

boundaries. These can be quite conspicuous especially at low bit rates.

(Qin et al) [13] proposes a post-processing method that can significantly reduce the tiling artefacts in low bit JPEG 2000 images.

These three types of artefact are all described as arising when encoding at low bit rates and comparatively high levels of compression. The levels of compression discussed in the previous section are much smaller than those discussed in the literature. These effects might in principle still be present even at low levels of compression; however they have not been detected in the compressed images produced in this work.

5. QUALITATIVE ASSESSMENT

5.1 Introduction

The preceding sections described a quantitative assessment which identified two best pairs of images for the top of range camera N05 and also from B07, one of the standard production cameras. The PSNR of the former pair was 43.00dB and the PSNR of the latter pair was 38.86dB. The degradation resulting from progressively greater compression was also assessed for both of these best match pairs of images. Within each of these series of compressed images two neighbouring images were identified: one with a PSNR value just greater, and one with a PSNR value just lower than the PSNR between the best match pair of master images.

We now report on a qualitative assessment which used a questionnaire, in which the relationship between the variations between these best match pairs and the neighbouring lossy compressed versions of one of the original master files formed a central part.

Two further types of assessment question were included in the questionnaire, regarding (a) the 'suitability for envisaged use' of a range of compressed images, and (b) a comparison between minimally lossless images and alternative lossless master images.

5.2 Questionnaire Design

There were three groups of questions, with ten questions overall.

The first group comprised four questions to compare the best match pair of original images against compressed versions of one image from each pair, as follows.

For device N05 one image of the best pair, N05B4, was designated as the original, and three alternative images were presented. Two of these were compressed lossy images and the other was the other lossless master file. The three images were the G3 and G4 lossy compressed images derived from N05B4 and master image N05B5.

Responses were sought indicating which image was least different and which image was most different from the original.

A very similar second question used a different small sample from the images from N05B5 and the G4 and G5 compressed versions from N05B4. The combined result of these two questions enabled us to relate the perceived difference between N05B4 and N05B5 to those from three lossy compressed versions of N05B4 – G3, G4 and G5.

Two further questions repeated this process with the other best match pair B07B3 and B07B5, relating the difference between them to those from three lossy compressed versions of B07B3 – namely G4, G5 and G6.

The second group of questions assessed the suitability of a range of compressed images for envisaged use. Image samples were taken from different types of content: a western manuscript, a music manuscript, and an eastern manuscript. The first two samples were presented at normal full resolution and the last at a magnification of 20. A lossless image was designated as the original and six alternative images were presented. One of these was the lossless original and the remaining five were the progressive more compressed lossy images G7 to G11 derived from the original. Responses were sought on whether the images were considered perfect, acceptable, marginal or unacceptable.

The third group of questions compared minimally lossless images with alternative lossless master images. The three questions again sought responses on whether the images were perfect, acceptable, marginal or unacceptable. The first image in this group was at normal full resolution and was based on the best overall set of images taken with device N05 with Item B. A sample taken from N05B4 was designated as the original and six alternatives were offered. These were the lossless original N05B4, a minimally lossless compressed version of N05B4, and lossless samples from N05B1-3 and N05B5.

The remaining questions in this group were both at magnification 60. One of these questions used samples from five alternative images: one was the original and the remaining four were all minimally lossless images derived in four different ways using kakadu. (These arise from whether a 'precise' flag is used during encoding, and separately if the same flag is used in decoding).

The final question was similar to the first question in the group except that this is at magnification 60. As before, a sample taken from N05B4 was designated as the original. There were six alternatives: the lossless original, a minimally lossless compressed version, and the other four were lossless samples from N05B1-3 and N05B5.

5.3 Questionnaire Responses

A survey questionnaire has been conducted comprising the ten questions described above. As responses were not mandatory their number varied. There were between 146 and 175 responses for each of the questions in the first group, and between 128 and 134 responses for each of the questions in the remaining groups. The conclusions are as follows:

The responses from the first group of questions were in line with the quantitative analysis.

Regarding the overall best match pair of images N05B4 and N05B5 the results from the questionnaire show that the difference between these images is comparable with the difference between N05B4 and its G5 lossy compressed version.

As noted in table 10 the PSNR comparing N05B4 and N05B5 is 43.00dB, and this lies between the PSNR values for G4 and G5 compression which are 44.48dB and 42.84dB respectively. These are relatively small differences in PSNR; hence the distinction between the images is not great and this does lead to a spread of responses:

- 90% indicate that the least overall change is from the lossy compressed version G3, and a further 4% with G4.
- 52% indicate that the most overall change is from N05B5, while 42% indicate that it is from G5.

Two questions have partial overlap when comparing only G4 and N05B5: with one question 72% indicated that G4 had least change while 21% indicated that N05B5 had least change, and with the

other question 21% indicated that G4 had most change while 74% indicated that N05B5 had most change

These lead to the conclusion that the difference in this best match pair is comparable with the difference between the original and the G5 lossy compressed version.

Regarding the overall best match pair of images B07B3 and B07B5 the results from the questionnaire show that the difference between these images is comparable with the differences between B07B3 and its G5 and G6 lossy compressed versions.

As noted in table 10 the comparison PSNR between B07B3 and B07B5 is 38.86dB, and this lies between the PSNR values for G5 and G6 compression which are 40.044 and 37.685 respectively. The responses were as follows:

- 96% indicate that the least overall change is from the lossy compressed version G4.
- 82% indicate that the most overall change is from G6 and 14% indicate that the most change is from B07B5.

Two questions have partial overlap when comparing only G5 and B07B5: with one question 71% indicate that G5 has less change than B07B5 while 26% indicated B07B5 had least change, and with the other question 64% indicated that B07B5 had most change while 33% indicated that G5 had most change.

These lead to the conclusion that the difference in this best match pair is comparable with the differences between the original and the G5 and G6 lossy compressed versions.

The questions in the second group were of similar structure except for a change in magnification.

At the original magnification compression of G8 or lower is typically considered perfect and G10 is considered acceptable or perfect. These correspond to compression ratios around 6 and 13 respectively.

However, at a magnification of 20 these drop to G2 and G3 respectively, where G2 is considered perfect and G3 is considered acceptable. These correspond to compression ratios around 1.8 and 2.3 respectively.

The questions in the third group investigated two different comparisons:

1. One pair of questions used the same master files, and included an original, a minimally lossless compressed version, and four lossless alternative master files. However, the two questions are at different magnifications.
2. The other pair of questions are both at magnification 60. One question comprises an original and different minimally lossless compressed images. The other question comprises an original, a minimally compressed image and four lossless alternative master images.

Regarding the first comparison:

- At the original magnification 84% considered the original to be perfect while 90% considered the minimally lossless version to be perfect. Between 4% and 5% considered that the alternative master files as perfect, 53% to 66% as acceptable, 23% to 30% as marginal, and 5% to 12% as unacceptable.
- At magnification 60 93% considered the original to be perfect, and 73% considered the minimally lossless version to be perfect. Between 1% and 3% considered that the alternative master files as perfect, 32% to 56% as acceptable, and 41% to 66% as marginal or even unacceptable.

Regarding the second comparison at magnification 60:

- 90% to 92% considered the minimally lossless versions to be perfect.
- Between 2% and 3% considered that an alternative master file was perfect, and the remaining 97% to 98% indicated that these were acceptable (32-56%), marginal (30-47%) or unacceptable (11-19%).

Minimally lossless images were deemed (mostly) to be perfect, whereas alternative master files were deemed to be only acceptable, marginal, or even unacceptable.

6. CONCLUSIONS AND DISCUSSION

6.1 Conclusions

Several conclusions may be drawn from the quantitative and qualitative assessments:

1. Images taken with even with a top of range camera show considerable variability despite all efforts to minimise difference in conditions. When compared with an original image at high magnification, alternative original images were considered merely acceptable, whereas at the same magnification, minimally lossless compressed vesrions of the original image were considered perfect.
2. The RMS variations in a minimally lossless compressed image are of the order of 30% - 40% of the variations between original lossless master images.
3. A minimally lossless file is typically 30-40% smaller than a lossless JPEG 2000 file.

Depending on a use case, which may be related to the type of content, an image may be compressed by a factor of between 3 and 6 compared with a lossless JPEG 2000 file and still be considered perfect at a magnification up to 20.

These conclusions, especially the recognition that there is considerable variability in the original images, question the need for images to be losslessly retained.

6.2 Discussion

The conclusions raise the question about how the value in an image arises. It could be associated with the image itself, perhaps because this image was taken by a famous person on a particular occasion. Or, more often, the value is in the subject of the image, such as a manuscript. Especially in the latter case, the conclusions suggest it might be appropriate to store the image in a slightly lossy compressed manner, especially if there is cost pressure on storage or transmission of the images.

Retaining images in a minimally lossless manner does reduce storage costs but appears to reduce the inherent value by a rather small, indeed we would argue negligible, amount. There may be concern that OCR may work less well and this should be investigated. However, today's OCR tools work with high quality images, and this paper has shown there is considerable variability in these. If OCR is compromised by minimally lossless compression then it would be highly likely that it would work with only a low proportion of quality digitized images. This clearly is not the case.

Depending on the type of content, for example with bulk digitisation, it would seem prudent to apply more compression

since a modest amount of compression can be applied without visual degradation of the image.

Some people express a belief that future tools will be developed that will reduce the noise in an image and thereby improve the quality in images that have already been taken. This could be done already if multiple images were taken of the same item, but this is not the standard process in today's cost efficient digitisation studios. There seems no greater reason to believe that the noise in single images of an item could in future be reduced better than the artefacts arising from slightly lossy compression – indeed the latter is arguably slightly more deterministic and therefore easier to tackle. Reliance on future improvements is thus questionable.

In terms of a business case, a baseline can be proposed based on the value and cost of a certain level of compression. An option can also be proposed to provide additional value but at additional cost by applying less compression but requiring more storage. Our results suggest that very little additional value is obtained in moving from minimally lossless to lossless but this would increase storage costs by roughly a half.

This work started with experimentation and over time it has helped establish a process for evaluation of noise in the digitisation process compared with the effects of lossy compression. Manufacturers are continually producing new camera models, so it would seem prudent to repeat these experiments periodically to provide a baseline assessment of the quality and repeatability of cameras as this changes over time.

7. ACKNOWLEDGEMENTS

We are grateful to Kjetil Iversen and his staff at the National Library of Norway and to Andrew Austin and his staff at the British Library for taking multiple images with several cameras and scanners. Ken Tsang developed software which analysed the sets of images.

8. REFERENCES

[1] Belbachir A. N. and Goebel P., "Medical Image Compression: Study of the influence of noise on the JPEG 2000 Compression performance," in 18th International Conference on Pattern Recognition, 2006.

[2] Chen L., Zhang X., Lin J. and Sha D., "Signal-to-noise ratio evaluation of a CCD camera," Optics and Laser Technology, vol. 41, pp. 574-579, 2009.

[3] Fang J. and Sun J., "Efficient Embedded Ringing Artifact Reduction for Low Bit-rate JPEG2000 Images," in Signal Processing, 8th International Conference , 2006.

[4] Fang J. and Sun J., "Ringing Artifact Reduction for JPEG2000 Images," in Proceedings Third International Conference on Intelligent Computing, ICIC 2007, Qingdao, China, 2007.

[5] Faraji H. and MacLean W. J., "CCD Noise Removal in Digital Images," IEEE Transactions Image Processing, pp. 2676-2685, 2006.

[6] Janesick J. R., Scientific Charge-Coupled Devices, Bellingham,WA: SPIE, 2001, pp. 605-719.

[7] Hashimoto M., Matsuo K. and Koike A., "JPEG2000 encoder for reducing tiling artifacts and accelerating the coding process," in International Conference on Image Processing, ICIP Proceedings., 2003.

[8] Kurosawa K., Kuroki K. and Akiba N., "Individual Camera Identification Using Correlation of Fixed Pattern Noise in Image Sensors," Journal Forensic Science, vol. 54, no. 3, pp. 639-641, May 2009.

[9] Liu C., Richard S., Kang S. B., Zitnick C. L. and Freeman W. T., "Automatic Estimation and Removal of Noise from a Single Image," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 30, no. 2, pp. 299-314, 2008.

[10] McHugh S., "Cambridge in Colour," [Online]. Available: http://www.cambridgeincolour.com/tutorials/image-noise.htm.

[11] Nasonov A. and Krylov A., "Scale-space method of image ringing estimation," in 16th IEEE International Conference Image Processing (ICIP), 2009; Pages: 2793 - 2796, 2009.

[12] Punchihewa A., "Objective evaluation of colour bleeding artefact due to image codecs," in 5th International Conference on Visual Information Engineering, VIE 2008. Pages: 801 - 806 , 2008.

[13] Qin X., Yan X.-L., Yang C.-P. and Ye Y., "Tiling artifact reduction for JPEG2000 image at low bit-rate," in IEEE International Conference on Multimedia and Expo, ICME '04. 2004; Pages: 1419 - 1422 Vol.2 , 2004.

An attempt at modeling differentiated storage for digitized collections: finding the balance between storage, costs and preservation of digitized publications[1]

Trudie Stoutjesdijk
Koninklijke Bibliotheek, Netherlands
Operations Department,
Collection Care Sub department
Trudie.Stoutjesdijk@kb.nl

ABSTRACT
The Koninklijke Bibliotheek (KB) digitizes the national collection of the Netherlands. Digitization leads to multiple versions of a publication: a digital access file, a digital master file, back-ups of the digital versions and the physical original publication. This in turn increases the need for storage capacity quickly. And raises questions like: Should all versions be stored? Do all the versions need to be preserved in order to ensure permanent access, and if so which ones should be preserved and how? Based on the collection care plan and the content strategy a differentiated storage policy is set up in order to establish a relation between the physical object and the digital counterpart(s). This method assigns value to different collection lots and is used to find out how to apply collection care in an efficient way.

Keywords
Storage policies; Collection Care; Permanent Access; Digitized collections

1. INTRODUCTION
As a national library the KB collects and maintains all publications that appear in the Netherlands, as well as a part of the international publications about the Netherlands. One of the large, labor-intensive challenges is to digitize all the books, periodicals and newspapers published in The Netherlands since 1470. Right now, nearly 10% (60 million pages) has been digitized. Digitization provides different versions of an object; therefore the amount of different versions of an object increases rapidly, as well as the storage costs. The most obvious approach, in order to reduce costs is to make sure that there are fewer copies of a publication. Questions are "Do we need to store all the versions of a publication? What representation of a publication is the object of preservation? Which ones do we want to remove, which one do we want to preserve? And which is most cost-effective method? " In order to answer these questions a close look at the current storage strategy is necessary. On the basis of the current collection care plan and archival storage system, a new storage model is proposed for digitized publications which distinguishes 5 different levels. Subsequently, there had been an investigation on potential costs savings and ways to use alternative solutions such as re-scan and conversion (or on-the–fly conversion) are possible.

2. KB MISSION
The mission of the Koninklijke Bibliotheek (KB), the National Library of The Netherlands, is to offer everyone everywhere access to all digital and printed publications that appear in the Netherlands. In addition, the KB fosters the establishment of a nev (digital) information infrastructure. Close cooperation between th KB, academic and public libraries is essential to grant everyone i the Netherlands access to scientific information. In order to achiev this goal, a transition from physical to digital is necessary.

3. THE BALANCE BETWEEN COLLECTION AND MANAGEMENT OF DIGITIZED PUBLICATIONS.
Our Collection Development Program [1] underpins the goal t make the KB collections digital. The program makes it clear whe to choose paper, and when to choose digital. It also explains th conditions of the strategy 'everything published in and about th Netherlands.' In 2003, KB's e-Depot became operational. It wa designed to preserve the electronic publications of Dutc publishers, in agreement with the Dutch voluntary deposi guidelines. Archival Agreements were signed with Dutch an internationally operating publishers. Ten years later, the e-Depo system, DIAS, is at the end of its natural life and a new digita preservation system (DPS) is being developed, called Digitaa Magazijn. The new DPS is a scalable digital archive; it consists o three major modules: (Workflow & Services; Process data an Metadata and Archival Storage) which represent the OAIS mode In 2012, the KB migrated collections from DIAS to the new DPS The next step is the development of new ingest workflows for th all the digitized collections and new born digital collections on th new DPS.

3.1 Collection development
The KB collects and preserves the printed and digital publication that are published in the Netherlands (e.g. the Netherland Collection), has an important collection of special old manuscript and early printed works, and a large number of digital database and e-journals. Our collections are of great importance as sourc material for (academic) research, as background reading fo university and professional education courses, and for everyon else who is interested in Dutch history, culture and society. Th selection strategy with regard to digital content is described in th Collection Plan 2010-2013. The transition from printe publications to a digital format is key priority. The KB wants t digitize in the coming years all books, newspapers and periodical which have been printed in the Netherlands since 1470. This is a endeavor that is beyond the capacity of the KB organization. W have therefore sought to cooperate, at first only with public parties but later also with private parties. Public partners are the Dutc House of Representatives, university libraries – in particular thos of Leiden and Amsterdam (University of Amsterdam) – and othe

[1] With many thanks to Tanja de Boer, Irene Hasslinger, Barbara Sierman en Marcel Ras.

cultural heritage institutions. In this way all the parliamentary papers and more than 10,000 Dutch early printed books from the end of the eighteenth century have been digitized. Various national and foreign cultural heritage institutions (archives, libraries) contribute to filling the website Historical Newspapers with nine million pages of newspapers dating from the seventeenth century to 1995.The KB sought cooperation with private parties for the first time. ProQuest is scanning our early printed books till 1700 and Google is digitizing our copyright free books from 1700 to around 1870.

3.2 Collection Care

Storage is one of the main costs of Collection Care. . In order to guarantee permanent access to the digital cultural heritage of the Netherlands we need to store our collections as efficient as possible. Our Collection Development Plan is complimented is complimented by our Collection Care Plan [2] that sets out a strategy for integrated, efficient and effective collection care for both digital and physical collections based on the following principles [3]:

- Integrated collection care for digital and physical objects
- Classification of collections into larger unities
- Valuation of collections
- Risk identification
- Different levels of collection care
- Care redirected from the most valuable collections, to those where the highest loss of value is indicated

Table 1: Values

primary criteria	secondary criteria
informational value	Use
aesthetic value	Completeness
historical value	Condition
social value	Provenance

It is neither possible nor necessary to apply the same care to all the physical and digital collections. Simply because there are differences between collections and the care they need. Not all collections are equally important nor are they equally vulnerable. The best care should go to the collections for which he greatest loss of value is expected. In order to be able to value the collections KB have divided physical and digital collections into lots or categories. There are 25 different lots: 14 lots in the digitized collections; 9 in the physical collections. These lots have been submitted to valuation by the collection specialists based on the defined values in table 1.[2]

After value and risk-assessment a set of preservation levels for the lots can be defined. The preservation levels determines the actions that are aimed at preventing loss of value as well as focusing on the loss of value for group of objects. The goal is to give just enough care to maintain the ability to retrieve, view online and use digital material in the face of rapidly changing technology.

Table 2: Classification levels

Preservation level	1. Lowest	2.	3.	4.	5. Highest
Representation available?					
-Digital Master	No	No	Master light	Preservation master	Preservatio master
- Access file	No	Yes	Yes	Yes	Yes
- Physical original	No	Yes	Yes	Yes	Yes
Preservation copy available?					
	No	No	Physical original	Preservation master	- Physical original - Preservatio master
Replacement by representation desirable?					
	N/A	Access file	Access file and Master light	Access file	Access file

The identification of values and risks to specific values will make i possible to determine the specific nature and amount of care for al the collections. Resources will be spend in a more effective an objective manner[3]. At the moment we are working on the fina value set of the lots. The emphasis in this paper is on th relationship between the physical original publication and th digitized counterpart(s). In anticipation of the outcome of the valu proposition, this model for physical storage of digitized collection is mainly based on the (secondary value) condition of the physica collections. The digitized collections are not under threat becaus they are managed in-house; the specifications are drawn up by th KB and the file formats are known (TIFF and JPEG2000).

3.3 Finding the balance

There are many aspects that play role in efficient and sustainabl storage of digitized publications. Digitization increases the amoun of different versions of an object rapidly. Digitized publication will yield a physical, digital master and access version besides th back-ups (2 times). That raises the question what level of storage i needed for the different versions. The level of storage is als determined by the desired degree of sustainability for the variou versions. And it is impossible to preserve all the versions at th highest preservation level and that will not be necessary. Findin the balance between these aspects is a real quest. Based on th collection care policy and the content strategy it is possible t establish a relation between the physical object and the digita counterpart(s) by assigning value to the different lots. This metho makes it possible to apply collection care in an efficient way. Ther will be a distinct relation between the state of the physical objec and the necessity of preservation imaging and sustainable storag of digital master files. A differentiated storage policy has bee applied on the digitized collections; this is based on:

- The availability of digital contents for the customer
- The vulnerability of the physical resources
- The sustainability of digital storage

[2] Based on the Australian publication *Significance*, published by the Heritage Collection Council in 2001. The digital version *Significance 2.0* was presented in 2009.

[3] Tanja de Boer and Matthijs van Otegem. *Moving to new digital storage migrating and reloading collections.* IFLA 2012.

3.4 Classification

Table 2 shows the classification of five levels, based on the values and the relationship between the different versions, the relationship between physical and digitized publications, the risks and the degree of effort that you want to apply to ensure permanent access of the collection(s). There is a distinction made between active and passive preservation; this only indicates which version is considered to be the master; that means the one that needs to be preserved for a long time. The concrete implementation of active and passive preservation effort needs to be completely based on the preservation policies.

Explanation of the different levels

Level 1: All these objects are available for use, the KB has no physical original and when usage drops, the subscription of the collection will be discontinued. This includes only licenses, there will be no conservation of objects in whatever form.

Level 2: This group is digitized to facilitate use. The main goal is the maintenance of the availability of the objects. The KB conserves the original objects passively: neatly stored on the shelf, whether or not compact stored in an air conditioned warehouse. The digital master need not be preserved. The digital derivative runs to the default backup and recovery procedure. This applies to all foreign titles in the Google project [4] (except if they still have value by particular provenance etc.).

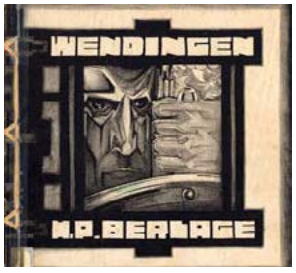


Figure 1: Magazine Wendingen

Level 3: This group contains objects that represent multiple values. The physical object is in a quite good condition and can be digitized repeatedly. That is why a digital master does not need to have the high quality of a preservation image; neither should it be preserved in an active way. A digital master light[5] [5] would do. In this way it could save costs of production and storage. The digital master will be preserved in a passive manner; it runs along in the usual backup and recovery procedures. The original physical object will be actively preserved if necessary, in order to keep the value as an object available. This type of object is in both the special collections (large parts of the 18th century collections) and in the Metamorfoze period[6] (e.g. art books and cultural important magazines as Wendingen and De Stijl). Customers are working basically with the digital version; the paper version is available for specific research questions.

Level 4: This group contains objects with high information value. Full reproduction by digitization is usually possible. In some cases, however, the material could be so fragile and easily subject to deterioration that digitization could only be done once, and it will not be possible to maintain the physical original. In this case the aim is to create a preservation image at very high quality. Thi preservation master will be the retention copy instead of th original physical. This occurs in case of the Metamorfoze perio where publications hardly represent value as an object.

No object is completely free from object value that is why the KI will not throw physical originals away. These objects will b conserved in a passive way: stored in an air conditioned warehouse Customers will have to work with the digital version and only i exceptional cases access to the paper original is allowed.

Level 5: Only a small part of the collection is so precious, fragil or difficult to digitize that it can only be digitized once. It follow that the quality of the digital master should be as high as possibl and maintenance is necessary, because there is no second chance t digitize. The physical object represents the primary values tha might not be reflected in the digital master: historical, aesthetic an / or society: for example, a bookbinding of William the Silent prints by H.N. Werkman (famous Dutch typographer, printer) Therefore the KB will preserve the original physical objec actively. The customer can use the digital copy, but has, in man cases also access to the physical object.

Figure 2: Bookbinding William the Silent

3.5 In summary

The collections at the first and second level are exclusively fo access. The first level exists of publications by subscription, the KB doesn't hold any objects only gives access to objects. Th second level focuses on digitization for access only. None of then need active preservation and only level two needs passiv preservation of both the original and the digital access file in orde to keep them accessible.

A large difference can be observed between level 2 and 3, th context and reference collection on the one hand and th Netherlands Collection on the other.

Level 3 to 5 will require sustainable access at high level, either b active conservation of the physical object (level 3) or the digita object (Level 4) or both (level 5).

4. DIGITIZED COLLECTIONS AND STORAGE COSTS

In order to discover how to guarantee permanent access to the KI collections as efficient as possible one must have a clea understanding of the costs. There are cost models that cover th entire preservation lifecycle, these are all useful models[7], bu there's still a strong development in the use of these models. One o the aspects of preservation is storage. In this model the focus wil be on the storage model in use by the KB. For the calculation o

[4] http://www.kb.nl/sites/default/files/docs/contract-google-kb.pdf

[5] The Master light digitalization quality level is intended for digitalizing originals whereby color accuracy is slightly less significant. Examples include books, newspapers, magazines and hand-written material.

[6] http://www.metamorfoze.nl/english/home

[7] Keeping Research Data Safe (KRDS); Cost Model for Digital Preservatior (CMDP); Digital Preservation for Libraries (DP4lib); Life Cycle Information for E-literature (LIFE3).

storage costs the KB uses a Total Cost of Ownership (TCO) for the entire storage infrastructure (including business and office storage costs[8]). Figure 1 shows the tiers and TCO of 2013 [4]; and the indicators for storage per TB. The storage costs per page are shown in table 3. The costs for digitization are based on a cost model and broken down by digital master files for permanent access and access files for current access. The following costs are distinguished: Capital costs, scanning costs and material costs. Based on this model and the production figures key performance indicators have been set (table 3).

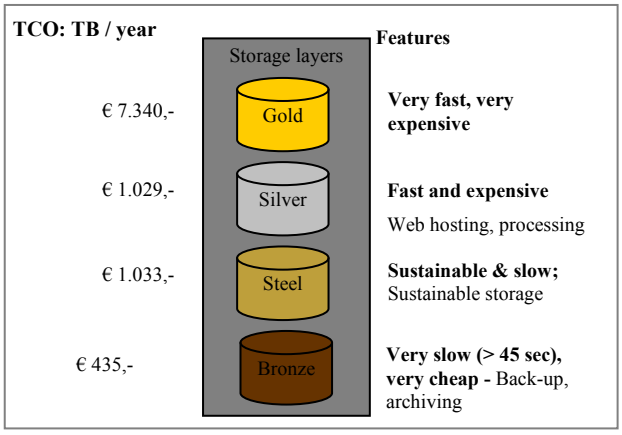


Figure 3: storage layers and costs

Cost Savings for storage of digitized publications

Currently the output of digitization process is a digital master and a digital access file. At the KB they are stored on different storage layers:

Silver: 1x digital access file
Steel: 1x digital master and in many cases 1x digital access file
Bronze: 2x back-up of tier steel (digital master and digital access file)

Table 3: Cost indicators

Type of publication	TB / page	Storage level / tier			Total costs / page	
Books		Bronze	Steel	Silver	Storage	Digitization
Master	0,00001	0,00435	0,01033		€ 0,01	€ 0,72
Access file	0,000001	0,000435		0,001029	€ 0,00	€ 0,56
Master & Access	0,000011	0,01914	0,01136	0,011319	€ 0,02	€ 1,28
Newspapers	TB/page	Bronze	Steel	Silver	Storage	Digitization
Master	0,00002	0,0087	0,02066		€ 0,02	€ 1,08
Access file	0,000006	0,00261		0,006174	€ 0,01	€ 0,93
Master & Access	0,000026	0,04524	0,02686	0,026754	€ 0,05	€ 2,01
Journals	TB/page	Bronze	Steel	Silver	Storage	Digitization
Master	0,000004	0,00174	0,00413		€ 0,00	€ 0,77
Access file	0,000001	0,000435		0,001029	€ 0,00	€ 0,61
Master & Access	0,000005	0,0087	0,00517	0,005145	€ 0,01	€ 1,38

[8] The components are partly based on the white paper "Four Principles for Reducing Total Cost of Ownership (2011 Hitachi).

At level 2 can yield cost savings because there will not be digital master files; this means that there are no production or storage costs for digital masters, only costs for digital access files. This could reduce the costs with 30 – 40%. At level 3 a digital master light will be created; a master light could require less image quality than a preservation master which could reduce the size of a digitized publication and lower costs of production and especially for storage. Digital master light criteria could be applied on objects of both the special collections (large parts of our 18th century collections) and in the Metamorfoze period (e.g. art books and magazines as Wendingen and De Stijl). But just now we do not have publications that are digitized conform the master light guidelines nor do we have cost indicators. As shown above, the application of the five level classification model reduce the storage costs of digitized publications.

4.1 Rescan and conversion to reduce storage costs

There can be several reasons for creating new digital master or digital access files. The access file no longer meets the requirements of the user, technologies offers new opportunities, possibly better and smaller digital masters or the original physical decay appears to be stronger than expected... Subsequently other additional methods to save costs were examined: rescanning and conversion.

Table 4: Rescan options

	Books		Newspapers		Journals	
Level 2: no master	storage	digitization	storage	digitization	storage	digitizatio n
Master & access	€ 0,04	€ 1,28	€ 0,10	€ 2,01	€ 0,02	€ 1,38
Access	€ 0,00	€ 0,56	€ 0,01	€ 0,93	€ 0,00	€ 0,61
Savings € / page	4%	30%	10%	32%	9%	31%

Rescanning

Rescanning, i.e. re-digitization of (parts of) the collection, of an object is a way to get a new digital representation. Rescanning is only possible if the original is present and in good physical condition. Again, assuming a classification of five levels of retention, based on the relationship between physical and digital, the possibility and desirability for any rescanning be determined.
For objects of Level 4 and 5 (publications with informative value and object value) rescanning is undesirable and sometimes even impossible. Only the digital master can serve as a source for new derivatives. Rescanning is a costly affair, whether rescanning is done to create an access or digital master file. Therfore clear criteria should be drawn up to decide in which exceptional cases rescan schould be done. These criteria should reflect the wishes of the customer, the physical condition of the original and technological developments. For the manufacture of a digital access file, re-conversion can offer a solution, in particular for vulnerable physical collections. Rescan on the basis of the known data is not a suitable tool to use in the KB-storage strategy. The development of conversion and conversion on-the-fly can avoid rescanning.

Conversion and/or on the fly conversion

In this article conversion refers to the method to derive a new an access file from the digital master. A large part of the digitized publications is stored in JPEG2000. One of the guiding principles to use JPEG2000 was the ability to reduce the overall storage

requirements by creating smaller files. The digital JPEG2000 master can serve as source for the access files. There are several ways to deal with coversion, one can create access files in advance and store them in the same way as the current storage of access files takes place, or ceate an access file at the time a publication is requested, on-the-fly. Conversion on the fly will reduce the storage costs and will directly benefit those who want to use the KB collections online. But on-the-fly conversion has other objections, it is a system intensive activity that could create a bottleneck in the delivery to the end user [9]

For the collections that are classified at level 4 and 5, conversion and "conversion-on-the-fly" could be an appropriate and efficient method for storage and permanent access of the publications. In these cases there is no reason for rescanning. Conversion of digital objects seems to offer a considerable advantage of saving cost on production and storage of the digital access files on the expensive tier silver. There only needs to be one derivative to be generated at the time at a customer's request. There is little experience with conversion or on-the-fly conversion from digital master files to digital access files. This technique has not been applied yet. It is advisable to do research to determine whether conversion can be used for preservation and mobilization purposes. Therefore the research department is asked to investigate the applicability of this technique fo the digitized publications.

5. CONCLUSION

In this article a model is developed to reduce the storage costs of digitized publications at the KB. The model reflects a balance between collecting of publications at large scale, management of them for access and long-term, and costs. Finding this balance is important to keep permanent access of the KB collection affordable. Based on the current collection care plan and archival storage system, we proposed a new storage model for digitized publications with 5 distinct levels. By using this model it became clear which publications to preserve and how to preserve them. Transparency of the costs tells how expensive digitization and storage of publications are. It also gains a clear understanding of possible cost saving alternatives: reduce redundancy (do not store Digital access files on steel and silver, nor 4x on bronze), the creation of new digital master and/or digital access files by rescanning or conversion.

Rescanning is not feasible for publications that are in vulnerabl state. Conversion might seem, from a cost efficiency point of viev preferable to that of rescanning. Investigation of the conversion on-the-fly conversion technique is necessary to gain insight into th benefits of this method. In particular with respect to applicability performance and efficiency.

6. REFERENCES

[1] Koninklijke Bibliotheek, Collectieplan 2010-2013: Fysiek en digital integraal. Available from http://www.kb.nl/en/organization-and-policy/collection-development-programme-2010-2013 (2009) accessed 8 March 2013.

[2] Koninklijke Bibliotheek. Collectiebehoudsplan 2010-2013: Fysiek en digitaal integraal. http://www.kb.nl/organisatie-en-beleid/collectiebehoudsplan-2010-2013 (2009) accessed 15 March 2013.

[3] Boer, Tanja de, and Otegem, Matthijs van, Moving to new digital storage: migrating and reloading collections. In *78th IFLA General Conference and Assembly* (Helsinki, 2012). http://conference.ifla.org/past/ifla78/102-boer-en.pdf

[4] Hoeven, Jeffrey van der, and Zavaros, Rogier, March 20, 2013. KB Kennissessie : Expert meeting TCO opslag. http://intranet/kb-breed/kennissessies/eerdere-kennissessies/expert-meeting-tco-opslag-20-maart-2013; accessed April 4 2013.

[5] Dormolen, Hans, January 2012. Metamorfoze Guidelines: image Quality, version 1.0 of January 2012. http://www.metamorfoze.nl/sites/metamorfoze/files/bestanden/ichtlijnen/Metamorfoze_Preservation_Imaging_Guidelines_1.0 pdf

[9] Knijff, Johan van der, at jpeg2000wellcomelibrary.blogspot.nl/

Preservation Aspects of a Curation-Oriented Thematic Aggregator

Dimitris Gavrilis
Digital Curation Unit
Athena Research Center
Artemidos 6 & Epidavrou
Maroussi, Greece
+30 2106875425, GR
d.gavrilis@dcu.gr

Stavros Angelis
Digital Curation Unit
Athena Research Center
Artemidos 6 & Epidavrou
Maroussi, Greece
+30 2106875425, GR
s.angelis@dcu.gr

Christos Papatheodorou
Dept. of Archives and Library Science,
Ionian University, Corfu, Greece and
Digital Curation Unit
Athena Research Center
Artemidos 6 & Epidavrou
Maroussi, Greece
+30 2106875425, GR
papatheodor@ionio.gr

Costis Dallas
Digital Curation Unit
Athena Research Center
Artemidos 6 & Epidavrou
Maroussi, Greece
+30 2106875425, GR
c.dallas@dcu.gr

Panos Constantopoulos
Digital Curation Unit
Athena Research Center
Artemidos 6 & Epidavrou
Maroussi, Greece
+30 2106875425, GR
p.constantopoulos@dcu.gr

ABSTRACT
The emergence of the European Digital Library (Europeana) foregrounds the need for aggregating content using smarter and more efficient ways taking into account its context and production circumstances. This paper presents the main functionalities of MoRe, a curation oriented aggregator that addresses digital preservation issues. MoRe combines aggregation, digital curation and preservation capabilities in a package that shields content providers from changes, and that ensures efficient, high volume metadata processing. It aggregates data from a wide community of archaeological content providers and integrates them to a common metadata schema. The system provides added-value digital curation services for metadata quality monitoring and enrichment so that to ensure metadata reliability. Furthermore it provides preservation workflows which guarantee effective record keeping of all transactions and the current status of the repository.

Categories and Subject Descriptors
H.3.7 [**Information Storage and Retrieval**]: Digital Libraries – *system issues*

D.2.12 [**Software Engineering**]: Interoperability

General Terms
Documentation, Performance, Design, Reliability, Standardization.

Keywords
Digital curation, aggregators, Europeana, CARARE, workflow, metadata enrichment, digital preservation, micro services.

1. INTRODUCTION
The emergence of the European Digital Library (Europeana) presents the need for aggregating content from multiple content providers and delivering this content to Europeana in a single metadata schema and in a uniform way. The CARARE project (Connecting Archaeology and Architecture in Europeana – http://www.carare.eu/) has delivered successfully over 2 million records (about 10% of Europeana's total content) from over 22 different content providers. The cultural assets made available are very diverse, from prehistoric and Iron Age archaeological survey results to complex Mediterranean archaeological sites and historic buildings. The digital resources representing such assets are also heterogeneous, ranging from paintings and prints to photographs, archaeological and architectural plans, sections and drawings, and, increasingly, digital 3D models.

The challenge faced by CARARE was that each content provider had their information in their native schema, using different ways to describe heterogeneous objects. Heritage assets are associated with geographic information, both in the form of geographic coordinates according to some grid standard, and in the form of named geographic entities such as historical place and area names; as expected, contentproviders used different coordinates systems, place names etc. Moreover archaeological sites are characterized by a nested mereological structure, being composed of buildings, each of which is also composed of particular architectural elements. Thus the main requirement for the descriptive metadata of such resources is to represent architectural and archaeological assets at quite different levels of complexity.

CARARE was the first of Europeana's projects to employ operationally the recently defined Europeana Data Model (EDM) [3]. EDM is a semantic graph schema that allows for a rich representation of a digital record. However it is still under development, so CARARE was facing the challenge of accommodating continuous changes in its delivery metadata schema. Additionally, even when EDM reaches a stable status, a part of the partners' metadata information content might be lost in the process of mapping to EDM, which is a generic schema and not especially suited to capture archaeological monument

documentation. An important challenge was, therefore, how to guarantee the preservation of the integrity of original archaeological information supplied by content providers.

This paper presents Monument Repository (MoRe), a repository system which addresses these issues by operating as an information broker between content providers and Europeana, offering value added curation services.

2. BACKGROUND

The traditional approach to aggregating metadata and links to digital resources into Europeana involves an aggregator [2] [11], which implements a crosswalk to transform original metadata records to records following a common output schema such as Europeana Semantic Elements (ESE) [4] or EDM. The crosswalks are based on a set of rules that map a source schema to the target schema (ESE or EDM).

CARARE represents a significant departure from this architecture. It introduces the notion of an information broker – an intermediate repository acting as a mediator – intended to ensure the integrity, authenticity and content enrichment of metadata provided to Europeana by heterogeneous collections. The overall architecture is shown in Figure 1. The content supplied by providers comprises administrative/scientific national registries of sites and monuments, archaeological museum collections, collections of 3D models describing any of these types of objects, as well as digital historical document collections such as the Visual Fortune of Pompeii archive. The metadata of all of these sources are transformed to a rich, thematic (in our case: sites and monuments) schema, the CARARE schema [9], and are stored in the CARARE repository, implemented using the Monument Repository system (MoRe) The CARARE schema is an application profile drawing on MIDAS Heritage, LIDO and the CIDOC CRM. The CARARE metadata are aggregated and delivered into the common format now used by Europeana to describe its content, the Europeana Data Model (EDM).

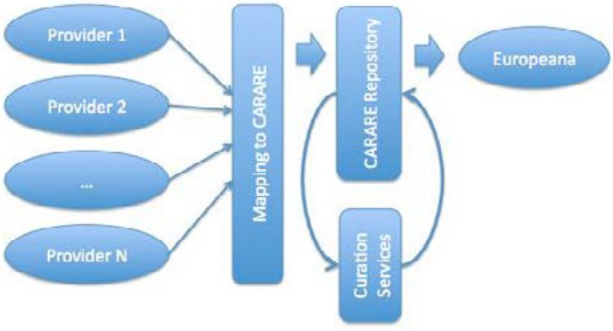


Figure 1. Overall architecture of CARARE

Compared with Europeana's approach for content aggregation, the repository approach followed by the CARARE project has the advantage of having the entire content available thus allowing to perform tasks at repository level, collection level, and content provider level, according to need. The architecture is based on a trusted repository approach [7] [8] and it aims to provide "reliable, long-term access to managed digital resources to its designated community, now and in the future" [12]. Hence MoRe demonstrates an organizational system that curates the archaeological information in accordance with commonly accepted standards and conventions.

This architecture matured alongside with the progress of the CARARE project, as one goal was for the repository to be flexible enough to tackle possible challenges that may appear. This allowed for the introduction of added value features, such as new services along the way of metadata harvesting. For instance, the usage of edm:Place element was introduced at a time when CARARE was already delivering content to Europeana. This element presented the need for visualization of information objects over a map. MoRe had to incorporate this element and provide the appropriate information to Europeana without necessitating a change in the original metadata or extra effort by content providers.

3. MONUMENT REPOSITORY (MoRe)

The mission of MoRe is to support the effective management of supplied information with minimal content providers' involvement. To this end it provides:

- versioning support for subsequent ingests of the same digital objects
- preservation services
- curation services

MoRe was built on top of Mopseus [5] [6], a Fedora-commons based digital repository developed by the Digital Curation Unit - Athena Research Centre.

3.1 Repository architecture

The repository architecture (Figure 2) consists of a core layer of services that receive information packages, pre-process them and store the metadata (datastreams) in a Fedora-Commons installation. The indexes of those datastreams are stored in a MySQL database. The metadata supplied by the content providers is transformed to the CARARE schema, stored, preserved, curated, and then made ready for publication.

MoRe functionalities are based on the implementation of micro-services, i.e., "small and well-defined procedures/functions that perform a certain task" [1] [13]. Chains of micro-services implement a larger macro-level functionality, which are called actions. Micro-services offer modularity in the construction of the MoRe workflows as a feature, and in tandem provide system administrators with full control of what happens in a workflow.

The core services of the repository, along with a set of curation services, have been developed in Java and run on an Apache Tomcat server. All services are orchestrated by a workflow engine and mainly operate at datastream level, although there are services that use only the MySQL index database. For instance, the clustering of records according to geographical proximity needs only to access the relevant indexes.

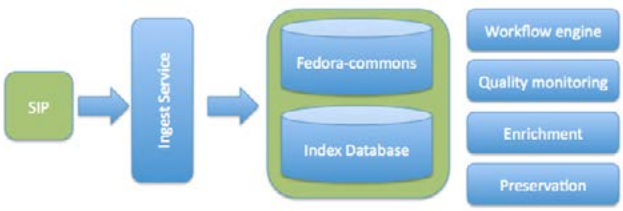


Figure 2. Monument Repository (MoRe) architecture

MoRe is fully OAIS compliant and handles three distinct types of information packages – Submission, Archival and Dissemination (SIP, AIP, DIP) - following certain specifications. Submission packages are created on ingestion and include the native (content provider's) metadata, the XSLT document that transforms the source metadata to the CARARE schema, as well as the corresponding CARARE metadata. All this information is accompanied by a technical metadata XML file (Figure 3) and ensures the tracing of provenance.

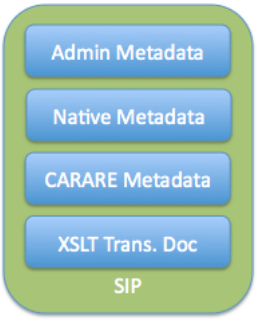


Figure 3. Submission Information Package

Each information object becoming part of the repository is wrapped into an Archival Information Package (AIP) which includes the SIP datastreams, as well as a PREMIS [10] datastream that contains a log with information about the ingestion of the object and relevant events (such as the datastream generation events, curation events, etc.), so that the object can be assigned preservation metadata. Finally, the Dissemination Package (DIP) includes both the CARARE and EDM datastreams. The EDM metadata (datastream) are harvested by Europeana, while the CARARE metadata are provided through the MoRe interface. The EDM datastream is created by the mapping service after the SIP package is ingested.

The services offered by MoRe are distinguished into core and curation services: the first are necessary to carry out the functions of content aggregation and delivery, while the second are intended for improving quality and adding value.

3.2 Core Services

Core services is the minimum set of services required in order to receive, transform and deliver content from the content providers to Europeana. These include:

3.2.1 Ingest

The ingest service is responsible for receiving submission information packages, performing various integrity checks and ingesting them into the repository. As SIP packages are received by the respective web service, they are stored in a temporary space awaiting to be verified. The verification process includes integrity checks on the SIP package in order to make sure that:

a) it contains the necessary information (e.g. package level admin and technical metadata);

b) all items in the package contain all the necessary XML datastreams (e.g. native metadata, CARARE metadata, admin metadata, XSLT transformation);

c) all XML documents are well formed;

d) each item contains valid item and provider identifiers.

After the verification step, the ingest service ingests the datastreams into Fedora-commons following the process below:

a) If the item (based on its provider identifier / native identifier) is new, a new digital object is created;

b) if the item exists, the existing identifier is retrieved;

c) all datastreams are ingested along with the corresponding PREMIS events;

d) the index service is triggered.

3.2.2 Indexing

Indexing is a fundamental service in all repositories. Mopseus comes with its own indexing mechanism which uses a descriptive XML document to define not only which parts of the metadata will be indexed, but also the structure of the SQL database that they will be indexed to. This approach simplifies and in part automates the work of other services such as the quality monitoring service. The repository manager is able to create the indexes and map them to any SQL schema. This approach allows to easily plug in services as they usually require specific table structures in the SQL database. For example, a service that discovers records that are in close proximity to each other, needs acces to a table that contains record identifiers and lat/lon coordinates.

3.2.3 Mapping

The CARARE Schema has been designed so as to capture the complexity of information represented within the CARARE aggregator, namely: collections, heritage objects, digital resources and activities. Thus, Archival Information Packages in the CARARE aggregator consist, in practice, of heterogeneous information, which needs to be re-expressed through mapping in order to allow harvesting and use by Europeana. All content ingested in MoRe is described in the CARARE Schema. Extracting these information objects to Europeana requires a mapping between CARARE Schema and EDM. This transformation is implemented through use of XSLT stylesheets. Depending on the native records, the transformation takes place on ingestion, or at a second step, if a particular set of data needs to be firstly de-duplicated (see Section 3.3.1). The mapping to EDM has been revised many times throughout the CARARE project, as the EDM Schema is still under development. Each time a new element was introduced or altered in EDM, the mapping had to be updated and the transformed objects reproduced and republished to Europeana. All this happens without the need of any effort on the part of content providers.

3.2.4 Delivery

The delivery service is responsible for delivering content through the OAI provider subsystem. The content to be delivered can be grouped per provider, per collection, per package (received package), and of course it has to take versioning into account (always the latest version is sent).

3.2.5 Repository Manager

The repository manager holds a key role in MoRe and in the CARARE project in general. The repository manager is in charge of executing second level checks on the data, making decisions about their overall quality and coordinating the proper operational scheme of the repository.

3.2.6 *Quality monitoring*

Quality monitoring is an essential part of an aggregator, as it informs content owners about the status of their information. It is based on policies, practices and performance that can be audited and measured in several ways, summarized per collection or even per submission package, as it is not feasible to inspect each information item separately. Some of the quality criteria are:

- Metadata completeness
- Unity of reference to information objects
- Element – Attribute completion
- Accuracy of spatial information

For example, metadata completeness measures whether the information captured per CARARE record meets the project's minimum acceptable standards. Although this task seems trivial at first glance, in a schema like CARARE it becomes somewhat more complicated. Consider the following examples:

a) Information completeness may vary among the top level elements of a CARARE object, and one could get a record with rich information in the heritageAsset and digitalResources elements, but minimal in the Activity elements. In this case, the overall quality is higher than estimated because the Activity element can be discarded during the mapping (to EDM) process.

b) Information can be captured in various ways. For example a spatial object may contain x/y coordinates without specifying the coordinate reference system, and these coordinates are not represented using WGS84. In this case, the actual quality is very low.

3.3 Curation Services

Curation services is a set of services running on MoRe, monitoring information objects and performing actions (curation actions) that aim to provide higher quality content. These are categorized in the table below and have various effects on the resulting metadata records.

Table 1. Curation actions

Action	Effect
Element & attribute cleaning	Homogeneity
De-duplication	Unity of reference, identification
Element & attribute fill	Improved completeness
Relation add	Additional information
Spatial transform	Homogeneity

For example, setting the language attributes (e.g. el, gre, GR to el) provides homogeneity to the resulting records and allows for building better services for end users. Spatial information is often encoded using different coordinate reference systems and has to be transformed to enable unified processing (for instance to the WGS84 system). In other instances, further information needs to be added to records, i.e.: a) a relation that denotes rights usage, b) a language attribute, or, c) the format type of a record.

The workflow used to execute curation services is especially important. Services that perform cleaning and simple element filling (with little built-in logic) are executed first. Following those are the more complex services which have more intelligent logic built into them, such as adding relations, performing de-duplication of records, etc. This sequence helps increase the information available to the more complex micro-services, thus yielding better results.

Below we discuss in some more detail two specific curation services, namely, De-Duplication and Geo-Spatial.

3.3.1 *The De-Duplication (De-Dup) curation service*

Each CARARE record may consist of four top-level elements: heritageAssets, digitalResources, collectionInformation, activities. The de-duplication service is responsible for:

- identifying duplicate top level elements among CARARE records of the same content provider / collection;
- removing the duplicates and replacing them with relations.

An illustration of how this service works is given in figure 4, where a set of 3 CARARE records are received (top row), and are transformed by the De-Duplication service (bottom row). Top level elements, such as the digitalResource of CARARE Record 2 and CARARE Record 3 (with id: D-1), are replaced by a relation (with id: H1) that points to the same element in CARARE Record 1. This process ensures unity of reference and identification, resulting in more robust sets of records.

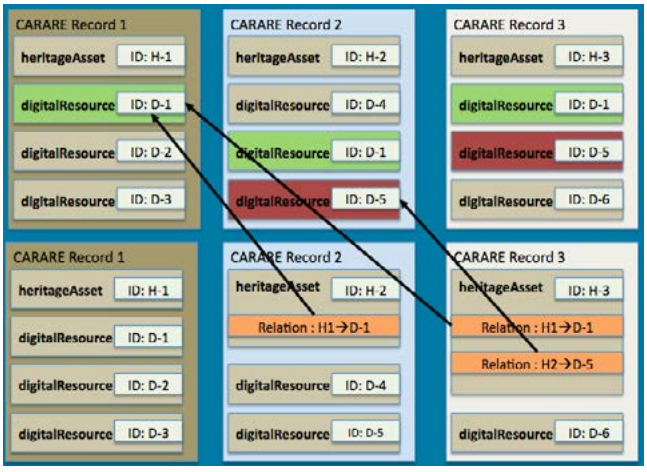


Figure 4. De-duplication example

3.3.2 *The Geo-Spatial curation service*

CARARE records descibe monuments. As a result, most of them contain spatial information in various forms: latitude/longitude coordinates (including coordinate system); historic place names address, and; country.

All the above information is encapsulated in a Spatial Element Block in CARARE schema [9], and usually not all information is provided. The Geo-Spatial curation service mainly performs three operations:

a) It checks coordinates (if provided) by verifying the correct coordinate reference system (Europeana only accepts WGS84), addressing errors in the provided x/y coordinates.

b) It checks the provided address, place name and other textual information and compiles them into one string (used in the target prefLabel element of the EDM set).

c) It geo-parses place names (if provided) using various openly available geo-parsers, and returns the place names they were mapped to the user. This feature is only provided to the end user through the UI.

An illustration of how the Geo-Spatial service works is given in Figure 5, where the spatial block from a CARARE record is displayed in the box on the left. This block of information contains a compilation of real use cases related to the geo-spatial information that had to be handled. Firstly, due to a mapping error, the x/y coordinates were concatenated in the x element. These are split (the parsing algorithm can detect and handle several cases). After the x/y coordinates are extracted, the coordinate reference system is checked. If it is different from WGS84, the coordinates are converted. If it is not provided, the x/y are checked to verify if they fall into the proper range. In the third step, the x/y are mapped to lat/lon, or vice versa (they must fall within the respective country). Finally, in the fourth step, the lat/lon are placed on a map and the country is checked out.

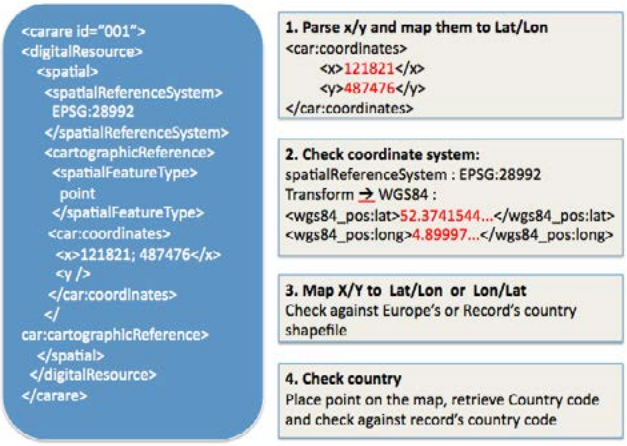


Figure 5. Geo-Spatial service example

3.3.3 Service Orchestration

When running a set of services in a streamline mode, the execution workflow is important especially with regard to preservation aspects. For example, in the execution workflow example presented in 3.4:

- In order to ensure integrity, the transform service must be executed after the Geo-Spatial service (because Geo-Spatial operates only on CARARE streams)
- In order to reduce processing resources required, the De-Dup service must precede all other services (because it results in fewer top level element sets requiring little or no processing).

Proper orchestration of these services helps reduce the amount of resources required, ensures integrity and helps formalize the overall ingest process.

3.4 Preservation Service

The preservation service is responsible for maintaining the metadata of the records provided to the repository, enabling their revision, versioning and validation, as well as maintaining the bond among various forms, and thus preserving provenance information. Each curation action that generates new content, or in any way modifies existing information, produces a new datastream version that is stored in Fedora-commons along with its PREMIS event log. A PREMIS [10] event log is maintained across the entire collection [5].

The Submission Information Package specification requires that each CARARE item is accompanied by its native record, the XSLT document that was used to transform between them, and the administrative metadata associated with the record. All these data are ingested as separate datastreams under the same item in MoRe, along with the appropriate PREMIS event (which is generated during ingest).

From a preservation point of view, the services layer of MoRe handles all preservation tasks. For example, consider a simple ingest of the 3 records shown in Figure 4, and assume that the De-Dup, Geo-Spatial and Mapping services are executed:

- **Ingest**. Each CARARE record is ingested
  - The Native datastream is added
  - The CARARE datastream is added
  - The XSLT (native→carare) is added
  - A PREMIS event record is generated and is added to the object
- **De-Dup**. For each record the De-Dup service processes
  - If the CARARE datastream is updated, a new datastream is added
  - A PREMIS event record is generated and is added to the object
- **Geo-Spatial**. For each record the Geo-Spatial service processes
  - If the CARARE datastream contains geo-information that needs to be updated, a new datastream is added
  - A PREMIS event record is generated and is added to the object
- **Mapping**. For each record the Mapping service processes
  - The CARARE record is transformed into EDM and the EDM datastream is added
  - A PREMIS event record is generated and is added to the object

This approach allows to track all the changes to the objects and to roll-back these changes if needed. Furthermore, the PREMIS records contain references to the services that operated on the datastreams, timestamps, user identifiers that possibly triggered the events, etc.

4. APPLICATION EXPERIENCE

During the 3-year CARARE project, over two million digital records were ingested, curated and delivered to Europeana using the system presented in this paper. From the 307 SIP packages that were received, 212 were ingested (the rest were discarded for not conforming to standards). These 212 packages contained approximately 3.6 million records from which only 2.6 million records were delivered to Europeana. The rest were discarded due to quality reasons, or were duplicates, a fact that demonstrates the importance of the De-Duplication service. The scale of digital records, as well as the number of the content providers accessing and making requests to MoRe, raised significant performance issues that were addressed successfully. For example, some typical big packages contained: 748.651, 487.882, 288.634 records. These had to be processed in short timeframes in order to

meet the strict deadlines that were laid out by the project. Using MoRe we were able to cope with the continuous changes in the EDM schema without having to burden content providers in order to re-harvest data. We also managed to provide clean, enriched records with the help of the curation services. Minimum amount of effort was required by content providers, as they had to provide their data once and all other processes were handled by the repository. This approach allows for future use of the same data without the need for further effort by content providers, as this data can be manipulated in the repository. Communication with content providers, including monitoring of original data quality and notification about issues with the data, was an important issue that was also successfully addressed.

5. CONCLUSIONS

This paper presented the added-value features of MoRe, a system that aims to aggregate, curate, preserve and make available quality metadata for archaeological monuments. MoRe aggregates information from a wide community of institutions and homogenizes it, obtaining interoperability between the diverse metadata schemas they use, on the basis of common well-documented policies and a common schema for metadata submission. In addition, MoRe provides procedures for access control and user authentication.

MoRe supports workflows for the effective record keeping of all transactions, as well as micro-services for the assessment of the completeness of the submitted metadata, combined with digital curation micro-services for the enrichment of aggregated metadata, and for increasing their quality and reliability. MoRe enables content curators and adminstrators to define workflows which implement policies for specifying how and at what level digital information is preserved, and how access is provided to users. It employs a functional de-duplication service, and ensures transformation to standardized geographic co-ordinates, both important features for accessing location-based, unique cultural heritage assets through an online user interface.

In conclusion, MoRe implements services that combine constitutive traits of both aggregators and trusted repositories. It offers a carefully prioritized workflow of services, optimized for high volume, industrial grade processing of complex metadata. It integrates curation services on top of established digital preservation standards, such as conformance with the OAIS model, and PREMIS metadata audit. It shields content providers from potential updates to the delivery schema. However, its most significant contribution is in empowering content providers to adopt good practices for the creation of digital materials, and to ensure the generation of clear, meaningful and homogeneous metadata for aggregation and online access.

6. REFERENCES

[1] Abrams, S., Kunze, J., Loy, D. 2010. An Emergent Micro-Services Approach to Digital Curation Infrastructure. *International Journal of Digital Curation*. 5, 1, 172 - 186. DOI= http://dx.doi.org/ 10.2218/ijdc.v5i1.151.

[2] Drosopoulos, N., Tzouvaras, V., Simou, N., Christaki, A., Stabenau, A., Pardalis, K., Xenikoudakis, F., Kollias, S. 2012. A Metadata Interoperability Platform. In *Proceedings of the Museums and the Web 2012* (San Diego, USA, April 11-14, 2012). MW2012. http://www.museumsandtheweb.com/mw2012/programs/a_metadata_interoperability_platform

[3] Europeana. 2012. Definition of the Europeana Data Model elements, version 5.2.3. http://pro.europeana.eu/documents/900548/bb6b51df-ad11-4a78-8d8a-44cc41810f22

[4] Europeana. 2012. Europeana Semantic Elements Specifications, version 3.4.1. http://pro.europeana.eu/documents/900548/dc80802e-6efb-4127-a98e-c27c95396d57

[5] Gavrilis, D., Angelis, S., Papatheodorou, C. 2010. Mopseus – A Digital Repository System with Semantically Enhanced Preservation Services. In *Proceedings of the 7th International Conference on Preservation of Digital Objects* (Vienna, Austria, September 2010). iPRES2010. 135-143.

[6] Gavrilis, D., Papatheodorou, C., Constantopoulos, P., Angelis, S. 2010. Mopseus – A Digital Library Management System Focused on Preservation. In *Proceedings of the 14th European Conference on Research and Advanced Technology for Digital Libraries* (Glasgow, UK, September 6-10, 2010). ECDL2010. Springer-Verlag, Berlin, LNCS 6273, 445-448.

[7] Jantz, R. 2005. Digital Preservation: Architecture and Technology for Trusted Digital Repositories. *D-Lib Magazine*, 11, 6 (June 2005). http://www.dlib.org/dlib/june05/jantz/06jantz.html

[8] Moore, R., Rajasekar, A., Marciano, R. 2007. Implementing Trusted Digital Repositories. In *Proceedings of the DigCCurr2007 International Symposium in Digital Curation* (Chapel Hill, North Carolina, USA, April, 2007). https://www.irods.org/pubs/DICE_DigcCur-Trusted-Rep-07.pdf

[9] Papatheodorou, C., Dallas, C., Ertmann-Christiansen, C., Fernie, K., Gavrilis, D., Masci, M.E., Constantopoulos, P., Angelis, S. A New Architecture and Approach to Asset Representation for Europeana Aggregation: The CARARE Way. 2011. In *Proceedings of the 5th International Conference on Metadata and Semantic Research* (Izmir, Turkey, October 12-14, 2011) MTSR 2011. Springer-Verlag, Berlin, CCIS 240, 412-423.

[10] PREMIS Presevation Metadata. http://www.loc.gov/standards/premis/

[11] Reis, D., Freire, N., Manguinhas, H., Pedrosa, G. 2009. REPOX – A Framework for Metadata Interchange. In *Proceedings of the 13th European Conference on Research and Advanced Technology for Digital Libraries* (Corfu, Greece, September 27- October 2, 2009). ECDL2009. Springer-Verlag, Berlin, LNCS 5714, 479-480.

[12] RLG - OCLC. 2002. *Trusted Digital Repositories: Attributes and Responsibilities*. Technical Report. RLG. http://www.oclc.org/content/dam/research/activities/trustedrep/repositories.pdf

[13] Ward, J.H., Wan, M., Schroeder, W., Rajasekar, A., de Torcy, A., Russell, T., Xu, H., Moore, R.W. 2011. *The integrated Rule-Oriented Data System (iRODS) Micro-service Workbook*, CreateSpace Independent Publishing Platform

Towards Concise Preservation by Managed Forgetting: Research Issues and Case Study

Nattiya Kanhabua
L3S Research Center
Leibniz Universität Hannover
Hannover, Germany
kanhabua@L3S.de

Claudia Niederée
L3S Research Center
Leibniz Universität Hannover
Hannover, Germany
niederee@L3S.de

Wolf Siberski
L3S Research Center
Leibniz Universität Hannover
Hannover, Germany
siberski@L3S.de

ABSTRACT

In human memory, forgetting plays a crucial role for focusing on important things and neglecting irrelevant details. In digital memories, the idea of systematic forgetting has found little attention, so far. At first glance, forgetting seems to contradict the purpose of archival and preservation. However, we are currently facing a tremendous growth in volumes of digital content. Thus, it becomes ever more important to focus, while forgetting irrelevant details, redundancies and noise. This holds true for better organizing the information space as well as in preservation management for making and revisiting decisions on what to keep. Therefore, we propose the introduction of the concept of *managed forgetting* as part of a joint information management and preservation management process in digital memories. Managed forgetting models resource selection as a function of attention and significance dynamics. Based on dynamic, multidimensional information value assessment it identifies information objects, e.g., documents or images of decreasing importance and/or topicality and triggers *forgetting actions*. Those actions include a variety of options, namely, aggregation and summarization, revised search and ranking behavior, elimination of redundancy, and finally, also deletion. In this paper, we present our vision for managed forgetting, discuss the challenges as well as our first ideas for its introduction, and present a case study for its motivation.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information filtering

General Terms

Human Factors, Measurement

Keywords

Digital Preservation; Dynamic Information Value Assessment; Time-aware Information Access; Managed Forgetting

1. INTRODUCTION

While preservation of digital content is now well established in memory institutions, such as, national libraries and archives, it is still in its infancy in most other organizations, and even more so for personal content. This is unsatisfying for two reasons: 1) with the growing volumes of and reliance on digital content there is a clear need for better long-term storage solutions in the organizational and in the personal context than the currently used backup strategies and 2) advanced and mature preservation technology is meanwhile available, also due to the intensive research and development work in this area in the recent years. For example, a variety of preservation platforms have been developed, such as, the SCAPE platform [25], which focuses on scalability or the platform developed in the PROTAGE project [11], which relies on a smart multi-agent architecture.

There are several obstacles for the wider adoption of preservation technology in organizational and personal information management: There is a considerable gap between active information use and preservation activities. Active information use refers to dealing with information objects for everyday private or professional activities, typically supported by some information management environment, such as, a content management system in an organization or a desktop environment in the context of personal information management. In addition, especially in personal information management, there is typically little awareness for preservation. Although the need for personal preservation has been recognized in theory [12, 14], this did not propagate to more practical settings and solutions yet. This is further aggravated by the fact that no benefits are seen for moving from more or less systematic backup to systematic preservation.

For improving preservation support in organizations, there is considerable research work underway as for example in the project ENSURE[1]. Lately, this also includes work on the preservation of business workflows [15]. In practical settings, systematic backups have become part of daily routine within organizations, at least with respect to a short- to mid-term perspective. However, the readiness to invest into preservation is low, if not enforced by legal regulations. Finally, establishing effective preservation and concise and usable archives still requires a lot of manual work for selecting content that is relevant for preservation and for keeping the archives accessible and meaningful in the long run, thus entailing expenses much larger than just the storage costs.

In this paper, we propose the introduction of the novel concept of *managed forgetting* as part of a joint information and preservation management process, in order to overcome some of the above obstacles. This concept is inspired by the important role of forgetting in the human brain, where forgetting enables us to focus on the things that are relevant instead of drowning in details by remembering everything. The idea of managed forgetting is to systematically deal with information that progressively ceases in im-

[1] `http://ensure-fp7-plone.fe.up.pt/site/`

portance and becomes redundant.At first glance, forgetting seems to contradict the idea of preservation, which is about keeping things, not about throwing them away. However, if no special actions are taken for long-term preservation, we already face a rather random digital forgetting process in the digital world today. This is triggered, e.g., by changing hardware, hard-disk crashes, or changes in employment. Furthermore, on a more global level there is a growing understanding that *forgetting* has to be considered as an alternative to the dominating keep it all paradigms, especially for information about individuals available in the Web [16].

We aim to replace such random forgetting processes with managed forgetting. In particular, we envision an idea of *gradual forgetting*, where complete digital forgetting is just the extreme and a wide range of different levels of condensation for preservation is foreseen. This concept is expected to both help in preservation decisions (also taking into account constraints for digital forgetting, e.g., legal regulations) and to create direct benefits for active information use by helping to keep the active information spaces more focused.

The rest of the paper is structured as follows: Section 2 describes the wider system context in which managed forgetting will be embedded in the ForgetIT project. Section 3 summarizes research challenges together with our first ideas for solving such challenges. Section 4 presents a case study in support of the motivation of managed forgetting. Finally, Section 5 concludes the paper with a description of the next steps towards realizing the concept of managed forgetting.

2. PROJECT AND SYSTEM CONTEXT

In our proposed approach, which will be implemented in the European project ForgetIT[2], our goal is to develop approaches and technologies for intelligent preservation management, which create a feasible and smooth path for preservation in the personal and organizational context and keeps the archived information concise, relevant and digestible by managed forgetting and contextualized remembering. For achieving its goal, the ForgetIT project will target: (a) enabling managed forgetting in information management and preservation management (cf. Section 3); (b) enabling contextualized remembering for keeping preserved content meaningful, useful and digestible through evolution-aware contextualization even when terminology, interpretation or context of use have changed considerably, and (c) closing the gap between information management and preservation management by introducing an approach for *synergetic preservation*. To validate the approach two application pilots will be built on top of the framework, one for preservation in the personal information management and the other in the organizational preservation management context.

2.1 Joint Model for Synergetic Preservation

For embedding the managed forgetting process we aim for an improved coupling between the information management system and the archival information system (AIS). This requires work on the conceptual level (preservation reference model extensions) as well as on the architectural level for system coupling. For institutional preservation organizations, reference models, such as, OAIS provide a solid foundation for the design and customization of preservation processes. Starting with ingest, OAIS describes very well how

[2]http://www.ForgetIT-project.eu

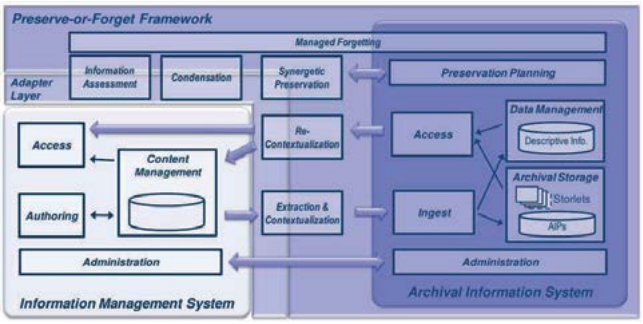


Figure 1: ForgetIT Architectural Approach.

content is transformed into self-contained archival packages and managed in the archival system. The part of the resource lifecycle which lies before ingest is, however, not included, although there is also work in the context of OAIS focusing on the pre-ingest phase such as the Producer-Archive Interface Methodology Abstract Standard (PAIMAS). Typically, this part of the resource cycle is described by information management workflows, covering tasks, roles, and resource states during the production process. To enable a tight connection between information management and preservation, these process models need to be coupled, to enable a seamless transfer of resources and their context information as well as to enable managed forgetting to be seamlessly applied.

It is planned to use OAIS as a starting point, and - taking into account other preservation process models as well (e.g., [23]) - and to develop a conceptual extension that covers the whole resource lifecycle. This reference model will treat issues such as when to create SIPs from resources of the active system for ingest by preservation storage, which context information from information management to preserve, or how to distribute responsibility for preservation tasks to information management roles.

2.2 Integration Architecture

In synergetic preservation, the roles of producer and consumer fall together. There may be other consumers, but one of the core consumers is the Information Management System. For the producer, preservation should be as transparent as possible; users which act as producers work in the active system and should not be forced to leave this environment for preserving their content. Consequently, the submission and access interfaces to the AIS should become part of the active system from the user's point of view. This poses an architectural challenge, because both information management as well as AIS come already with their own full-fledged software architectures. The aim is to achieve a tight integration without re-inventing a new integrated framework from scratch. The approach here is to use existing preservation architectures, and to realize the integration with an information system specific adaptation layer (see Figure 1). This layer connects system-specific content models, events, and processes to the corresponding generic preservation concepts implemented as part of what we call the Preserve-or-Forget framework, and the implementation of the managed forgetting process will be part of this framework.

A core factor for synergetic preservation is the smooth transition of content from information management storage to preservation storage. In addition, it is important to also support the reverse direction, i.e., to put the resources deliv-

ered by the preservation store back into active use. Depending on how far gone the information object is in its state of "inactivity", the object might be extracted into a format that is directly able to become ingested back into the active system, or to a format that is more platform independent and less likely to be directly ingestible in its original system. When re-activating a previously archived object, contextual links need to be re-created and/or updated to account for semantic shift (re-contextualization).

3. CHALLENGES

The introduction of managed forgetting into digital memories is a challenging task and its adequate combination with the goals of preservation has to be carefully investigated implying three key challenges:

- An interdisciplinary concept for flexible and gradual **managed forgetting** that meets **human expectations** and is driven by the goal of the digital memory complementing human memory;
- Development of flexible and multifaceted **information value assessment** methods in support of managed forgetting and in support of resource selection for preservation;
- Development of adequate **forgetting actions** especially for **quality-aware consolidation and concentration** for textual and multimedia content, such as, summarization, aggregation, detection of redundancy, and consideration of diversity.

3.1 Challenge: Meeting Human Expectations

Relevant State of the Art: In the field of psychology, aforementioned works [18, 27, 29] conducted subjective studies in order to shed light on understanding human remembering and forgetting. This can benefit digital preservation methods that aim at complementing the human ability to remember or forget information. From the Human-Computer Interaction (HCI) perspective, works related to digital preservation are, e.g., [4, 7, 8], which focus on system design for supporting the reminiscence of past events.

First Ideas in ForgetIT: Supporting managed forgetting in a digital memory is a novel concept, for which no former experience and best practices exist. It is therefore important to thoroughly analyze the human expectation for this process. An interdisciplinary approach is planned for this purpose. The idea is to investigate, what we can learn from the way a human memory forgets and remembers. Humans are, for example, very effective in (a) rapidly extracting the general gist of an experience, while forgetting many details, in (b) extracting common pattern of similar experiences avoiding the redundant "storage" of such pattern, and in (c) identifying data that are only temporally required and can be forgotten after task completion. Those and further characteristics of human forgetting will be further investigated. Selected characteristics will flow into a model for managed forgetting. The goal is, however, to complement not to copy or replace human memory. This perspective will create the highest benefit in the interaction of humans with digital memory. For analyzing the expectations towards managed forgetting user studies will be performed.

A further important source of inspiration for tailoring the managed forgetting process are the best practices and guidelines, which are already used in libraries and archives for selecting material for retention, transfer and destruction.

3.2 Challenge: Multifaceted Information Value Assessment

Relevant State of the Art: Forgetting basics [1, 9, 10, 22] are based on a decay theory, and an interference theory. There have been some works on modeling a temporal decay function, for example, applied to data streams [19] and exploited in information retrieval [13]. A recent work [20] considers different temporal document priors inspired by retention functions [17] considered in cognitive psychology that are used to model the decay of memory.

First Ideas in ForgetIT: Assessing the information objects in digital memory provides the basis for triggering managed forgetting actions, such as, condensation, contextualization and transition to the archive. We define two complementing information assessment values: *memory buoyancy* and *preservation value*. Memory buoyancy is inspired by the metaphor of information objects sinking down in the digital memory with decreasing importance, usage, etc., which increases their distance to the user. Memory buoyancy is influenced by a variety of factors in the following categories: usage parameters (such as, frequency and recency of use, user ratings, recurrent pattern), type and provenance parameters (information object type, source/creator) and context parameters (such as, relevance of resources as background information, general importance of topic, external constraints), and temporal parameters (age, lifetime specifications). The preservation value reflects the importance that the considered object gets preserved and will be used to decide if and when to archive an information object. Partly, the preservation value is influenced by similar factors as memory buoyancy, but it serves a different purpose: An object with a high value of memory buoyancy might already be moved to the archive (as a copy), because it has a very high preservation value, while staying still in direct uncondensed access to the user; an information object with low memory buoyancy and low preservation value might be preserved only in its condensed version or it might be decided not to preserve it at all. In this activity various factors influencing memory buoyancy and preservation value will be investigated as well as approaches for learning most effective factor combinations. Furthermore, approaches for enabling the user to explicitly and implicitly influence the values for memory buoyancy and preservation value will be developed, e.g., explicit expiry dates and lifetime specifications or tagging objects as non-forgettable.

3.3 Challenge: Flexible Forgetting Actions

Relevant State of the Art: Relevant research areas to forgetting actions for quality-aware consolidation include document summarization, duplicate detection, and diversity analysis. Automatic document summarization [26] is aimed at extracting the semantic content from a document in order to produce a well-formed and grammatical summary of what the document or document set is about and what its broad content is. Aforementioned works on detecting duplicate or near-duplicate documents has been mainly focused on different similarity metrics [5, 6, 28]. In the area of information retrieval, there is an interplay between redundancy, diversity and interdependent document relevance [3, 21].

First Ideas in ForgetIT: There are several forms of forgetting that will be supported including: changing the ranking of the "forgotten" object in a result list or not showing it as a result at all, replacing the object by a summary

object, marking the object as a deletion candidate etc. As an extreme the process will also support deletion as a forgetting option. Furthermore, managed forgetting will be used in several places of the information and preservation lifecycle: for focusing the content in active use, for helping in preservation decisions and for revisiting preservation decisions within the archive (gradual forgetting). Clearly there will be no one size fits all for managed forgetting, either. It is planned to define an adaptable framework for the managed forgetting process, which fixes the principle mechanisms of the process and can be customized along different dimensions: the parameters that are used for information assessment, the threshold used for memory buoyancy and preservation value for triggering forgetting actions and the options of forgetting considered. We will also investigate the use of a policy framework that supports the definition of different forgetting policies. Policies have been shown to be an intuitive and powerful tool in the area of security management, e.g., for specification of access rights. In the preservation context, besides customizing the forgetting process, policies also can capture external constraints, such as legal preservation requirements or business requirements (e.g., to make sure that information pertinent to obsolete product versions is preserved). Furthermore, we will also investigate into methods for detecting redundancies and for condensating textual as well as multimedia information objects.

4. DELETION BEHAVIOR IN ONLINE SOCIAL BOOKMARKING: A CASE STUDY

To support our motivation of systematic forgetting, we conducted a case study of analyzing deletion behavior in Online Social Bookmark and Publication Management System - BibSonomy [2]. The web-based system supports team-oriented publication management and social bookmark sharing. BibSonomy offers users an ability to categorize and archive two types of resources, i.e., *bookmarks* and *literature references*. In particular, a user can upload and share a resource, or label them with arbitrary words, so-called *tags*. In addition, an uploaded resource can also be deleted from the system by its owner when needed.

A formal model for BibSonomy is given as follows: U, T, and R are finite sets, whose elements are called users, tags and resources, respectively. Y is a ternary relation between them, i.e., $Y \subseteq U \times T \times R$, whose elements are called tag assignments, and the set P of all posts is defined as $P = \{(u, S, r) | u \in U, r \in R, S = T(u, r), S \neq \emptyset\}$ where, for all $u \in U$ and $r \in R$, $T(u, r) = \{t \in T | (u, t, r) \in Y\}$ denotes all tags the user u assigned to the resource r. The principal unit of our analysis is a post p, which is a transaction made when inserting a resource to the system. Based on the BibSonomy data model described in [2], there can be more than one transaction records associated to a resource uploaded. This is because a transaction record will be created for *each tag* assigned to the inserted resource. In this study, we do not leverage user tag information, and all transaction records belonging to the same resource ID will be regarded as one unit of study, or a *post* in our case. Thus, a post p is defined as a tuple $(u, r, time(r))$, where a user u is the owner of a resource r uploaded at *time(r)*.

In order to motivate the concept of managed forgetting, we investigate deletion processes manually performed by users over time, so-called *deletion behavior*. We obtained the publicly-available data dumps of BibSonomy consisting

Table 1: Statistics of distinct posts per user.

Type	Max	Avg	Std
All	119,678	370.87	2872.39
Bookmark	58,144	171.91	1292.09
Bibtex	119,678	198.96	2556.16

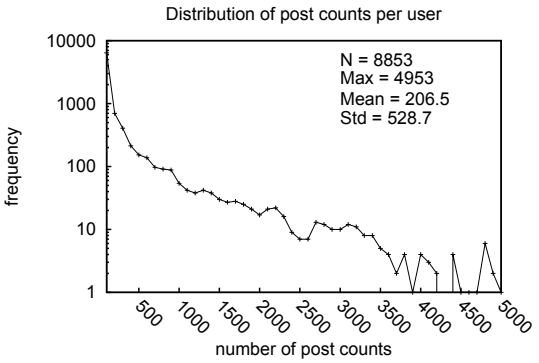


Figure 2: Distribution of post counts per user.

of 15 data snapshots, i.e., 2006-06-30, 2006-12-31,..., 2012-01-01, 2012-07-01, 2013-01-01, where the average time distance between any two snapshots is approximately 6 months. The dataset does not contain information about user names and demographics, thus our analysis was studied unobtrusively and compiled anonymised client based Web transaction logs. As of 1 January 2013, the number of active users in this study is 8,928 users, and basic statistics about distinct posts per user are shown in Table 1. The maximum numbers of bookmarks and bibtex posted per user are 58,144 and 119,678, respectively. On average, there is about 370 resources posted per user and the average of bookmarks and bibtex posted per user are 171 and 198, respectively.

As mentioned in [2], there are non-human users that automatically insert posts, e.g., the DBLP computer science bibliography.Therefore, we ignored such users with more than 5,000 posts from the analysis. To this end, we conducted the study in total of 8,853 users. Figure 2 shows the distribution of the number of distinct resources (post counts) per user. We conducted a detailed analysis by dividing users with respect to the number of their resources posted in total, into three groups: Group1 (10-100 posts), Group2 (101-1,000 posts), and Group3 (> 1,000 posts). Our hypothesis is that different groups of users can shed light on the different characteristics of deletion behavior among users who share posts from very few to very many.

Deletion behavior was studied by computing the number of posts added or removed made by each user at different time snapshots. For a given user u, the number of posts *added* at a particular time snapshot t_i can be computed as the *difference of two sets*, namely, the set of posts at current time t_i and the set of posts at the previous time snapshot t_{i-1}: $add(u, t_i) = P(u, t_i) - P(u, t_{i-1})$, where the type of post $p \in P$ can be either a bookmark or a publication reference (denoted *bibtex*). On the contrary, the number of posts *removed* at a particular time snapshot t_i can be computed as the difference of the set of posts at the previous time snapshot t_{i-1} and the set of posts at current time t_i: $remove(u, t_i) = P(u, t_{i-1}) - P(u, t_i)$.

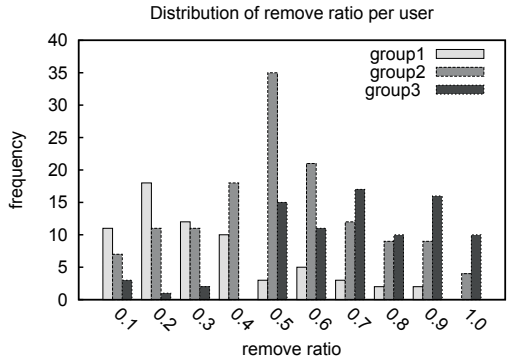


Figure 4: Remove ratios among different groups.

The trend over time of posts added or removed on average among three different groups is illustrated in Figure 3. In general, the results exhibit highly similar trends among different groups. Our observation is that, at each time snapshot, the number of added posts is greater than the number of posts removed in most cases, for all groups. This results in the increasing number of posts accumulated over time. For Group1 and Group2, the number of posts of the type *publication* is slightly higher than *bookmark*. It can suggest that users in the first two groups mostly share publication references than bookmarks, whereas the number of bookmarks posted by Group3 users are significantly higher than publication references.

In addition to raw counts, we also computed *remove ratio* as a fraction of the number of time snapshots a user deleted at least one post. For example, a user u has been a member since 2006, and the user deletion activity is observed 10 times during 15 snapshots in time. Thus, *remove ratio*(u) equals to $0.67 = (10/15)$. Figure 4 illustrates the distribution of users' remove ratios among different groups. The results show that the group of users with fewer posts (Group1) has fewer deletion activities, while the group with more posts (Group3) tends to delete more often.

What trigger a deletion process? Does the number of current posts or that of newly added posts influence the deletion? We sought to answer such questions by performing a correlation analysis by correlating: 1) deletion activities with the total number of posts (or bookmarks or bibtex) and 2) deletion activities with the number of *added* posts (or bookmarks or bibtex). Note that, we only considered any user u with *remove ratio*(u) ≥ 0.5. Table 2 shows the correlation results of deletion activities over time with the total number of posts (**Post**), the number of *added* posts (**Post+**), the total number of bookmarks (**Bookmark**), the number of *added* bookmarks (**Bookmark+**), the total number of bibtex (**Bibtex**), and the number of *added* bibtex (**Bibtex+**), respectively. In general, it can be observed that deletion is highly correlated with the number of resources added, but not the number of total resources users currently possess. Finally, Group1 shows highest correlation results between deletion and added resources in most cases.

Our final analysis is to determine whether given resources are still accessible online. This is motivated by the recent study *Losing my revolution: how many resources shared on social media have been lost?* by SalahEldeen and Nelson [24]. The work has estimated that 27% of resources shared in social media are lost and not archived after 2.5 years. Table 3 shows the total numbers and percentage of resources that were accessible online using their URLs (retrieved on 28 April 2013). On average, there are less than 83% of bookmarks and less than 74% of publication references that were still accessible. This observation suggests that it is important to automatically identify unavailable resources and trigger a forgetting action, e.g., tagging objects as forgettable or deletion, in order to help user handle obsolete information.

Table 3: Resources accessible on 28 April 2013.

	#Bookmark (%)	#Bibtex (%)
Group1	715 (87.73%)	546 (78.56%)
Group2	5,074 (81.34%)	4,396 (73.39%)
Group3	24,909 (78.21%)	3,984 (69.48%)

5. CONCLUSIONS AND FUTURE WORK

In this paper, we presented our vision for integrating the concept of managed forgetting into a joint information and preservation management process. This work is still in a very early phase. Nevertheless, we wanted to take the opportunity to discuss the idea of managed forgetting in the preservation community. As a consequence there is still a rich set of future work ahead of us, including: Foundations for the managed forgetting process building upon interdisciplinary work with cognitive psychology; a substantiated information value assessment model in support of the information value dimensions memory buoyancy and preservation value. This also includes the identification of the set of measurable parameters best to be used for estimating those values; and experiments for better understanding the constituents and mechanisms of managed forgetting, e.g., interactions with photo collections, and revisiting behaviors for Web users as well as organizational information seekers.

Acknowledgments This work was partially funded by the European Commission Seventh Framework Program under grant agreement No.600826 for the ForgetIT project (FP7 / 2013-2016).

6. REFERENCES

[1] P. Barrouillet, G. Plancher, A. Guida, and V. Camos. Forgetting at short term: When do event-based interference and temporal factors have an effect? *Acta psychologica*, 142(2):155–67, Feb. 2013.

[2] D. Benz, A. Hotho, R. Jäschke, B. Krause, F. Mitzlaff, C. Schmitz, and G. Stumme. The social bookmark and publication management system BibSonomy. *The VLDB Journal*, 19(6):849–875, Dec. 2010.

[3] Y. Bernstein and J. Zobel. Redundant documents and search effectiveness. In *Proceedings of the 14th ACM international conference on Information and knowledge management*, CIKM '05, pages 736–743, 2005.

[4] S. Bowen and D. Petrelli. Remembering today tomorrow: Exploring the human-centred design of digital mementos. *International Journal of Human-Computer Studies*, 69(5):324–337, May 2011.

[5] A. Z. Broder. Identifying and filtering near-duplicate documents. In *Proceedings of the 11th Annual Symposium on Combinatorial Pattern Matching*, COM '00, pages 1–10, 2000.

[6] A. Chowdhury, O. Frieder, D. Grossman, and M. C. McCabe. Collection statistics for fast duplicate document detection. *ACM Trans. Inf. Syst.*, 20(2):171–191, Apr. 2002.

[7] D. Cosley, V. S. Sosik, J. Schultz, S. T. Peesapati, and S. Lee. Experiences With Designing Tools for Everyday Reminiscing. *Human–Computer Interaction*, 27(1-2):175–198, 2012.

[8] M. Crete-Nishihata, R. M. Baecker, M. Massimi, D. Ptak, R. Campigotto, L. D. Kaufman, A. M. Brickman, G. R. Turner, J. R. Steinerman, and S. E. Black. Reconstructing the Past:

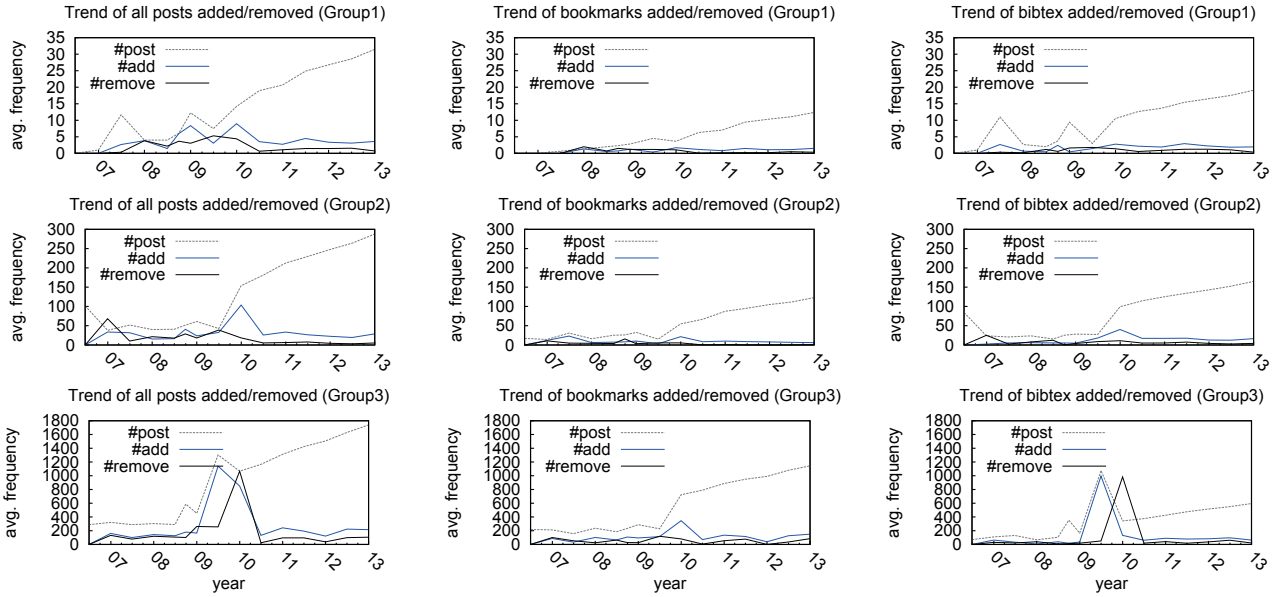


Figure 3: Trend over time of posts added/removed by user.

Table 2: Correlation of deletion behavior with the number of posts, bookmarks and bibtex.

	#Users	Post	Post+	Bookmark	Bookmark+	Bibtex	Bibtex+
Group1	13(66)	0.2632	0.5924	0.0875	0.3214	0.3413	0.6994
Group2	65(137)	0.2140	0.4247	0.1274	0.4195	0.3503	0.6249
Group3	73(85)	0.0918	0.3183	0.0845	0.3615	0.1591	0.4793

Personal Memory Technologies Are Not Just Personal and Not Just for Memory. *Human–Computer Interaction*, 27(1-2):92–123, 2012.

[9] H. Ebbinghaus. *Memory: A contribution to experimental psychology*. Teachers college, Columbia university, 1913.

[10] A. Heathcote, S. Brown, and D. J. K. Mewhort. The power law repealed: The case for an exponential law of practice. *Psychonomic Bulletin & Review*, 7:185–207, 2000.

[11] X. Jin, J. Jiang, and J. L. de la Rosa. Protage: Long-term digital preservation based on intelligent agents and web services. *ERCIM News*, 80:15–16, January 2010.

[12] L. Johnston. We are all digital archivists: Encouraging personal digital archiving and citizen archiving. In *Proceedings of iPres 2011 - 8th International Conference on Preservation of Digital Objects, Signapore, November 2011*, 2011.

[13] X. Li and W. B. Croft. Time-based language models. In *Proceedings of the twelfth international conference on Information and knowledge management*, CIKM '03, pages 469–475, 2003.

[14] C. C. Marshall. Challenges and opportunities for personal digital archiving. In C. A. Lee, editor, *I, Digital: Personal Collections in the Digital Era*, pages 90–114. Society of American Archivists, 2011.

[15] R. Mayer, S. Pröll, and A. Rauber. On the applicability of workflow management systems for the preservation of business processes. In *Proceedings of iPres 2012 - 9th International Conference on Preservation of Digital Objects, Toronto, Canada, October 2012*, 2012.

[16] V. Mayer-Schönberger. *Delete - The Virtue of Forgetting in the Digital Age*. Morgan Kaufmann Publishers, 2009.

[17] M. Meeter, J. M. J. Murre, and S. M. J. Janssen. Remembering the news: Modeling retention data from a study with 14,000 participants. 33(5):793–810, 2005.

[18] L. Mickes, T. M. Seale-Carlisle, and J. T. Wixted. Rethinking familiarity: Remember/Know judgments in free recall. *Journal of Memory and Language*, 68(4):333–349, May 2013.

[19] T. Palpanas, M. Vlachos, E. Keogh, D. Gunopulos, and W. Truppel. Online amnesic approximation of streaming time series. In *Proceedings of the 20th International Conference on Data Engineering*, ICDE '04, pages 338–349, 2004.

[20] M.-H. Peetz and M. de Rijke. Cognitive temporal document priors. In *Proceedings of the 35th European conference on Advances in Information Retrieval*, ECIR'13, pages 318–330, 2013.

[21] F. Radlinski, P. N. Bennett, B. Carterette, and T. Joachims. Redundancy, diversity and interdependent document relevance. *SIGIR Forum*, 43(2):46–52, Dec. 2009.

[22] D. C. Rubin, S. Hinton, and A. Wenzel. The precise time course of retention. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 25(5):1161–1176, 1999.

[23] M. Runardotter, H. Quisbert, J. Nilsson, A. Hägerfors, and A. Mirijamdotter. The information life cycle : issues in long-term digital preservation. *Arkiv, samhälle och forskning*, 1(1):17–29, 2006.

[24] H. M. SalahEldeen and M. L. Nelson. Losing my revolution: how many resources shared on social media have been lost? In *Proceedings of the Second international conference on Theory and Practice of Digital Libraries*, TPDL'12, pages 125–137, 2012.

[25] R. Schmidt. An architectural overview of the scape preservation platform. In *Proceedings of iPres 2012 - 9th International Conference on Preservation of Digital Objects, Toronto, Canada, October 2012*, 2012.

[26] K. Spärck Jones. Automatic summarising: The state of the art. *Inf. Process. Manage.*, 43(6):1449–1481, Nov. 2007.

[27] J. A. Sumner, S. Mineka, R. E. Zinbarg, M. G. Craske, S. Vrshek-Schallhorn, and A. Epstein. Examining the long-term stability of overgeneral autobiographical memory. *Memory*, Feb. 2013.

[28] M. Theobald, J. Siddharth, and A. Paepcke. Spotsigs: robust and efficient near duplicate detection in large web collections. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '08, pages 563–570, 2008.

[29] N. Unsworth, B. D. McMillan, G. A. Brewer, and G. J. Spillers. Individual differences in everyday retrospective memory failures. *Journal of Applied Research in Memory and Cognition*, 2(1):7–13, Mar. 2013.

Acquiring and providing access to historical web collections

Daniel Gomes, David Cruz, João Miranda, Miguel Costa, Simão Fontes
Foundation for National Scientific Computing
Av. Brasil, 101
1700-066 Lisboa, Portugal
{daniel.gomes, david.cruz, joao.miranda, miguel.costa, simao.fontes}@fccn.pt

ABSTRACT

Every day, unique valuable information that describes our current days disappears from the web. National archives or libraries have been keeping cultural heritage for centuries by collecting and preserving past generation objects or printed media. Now, it is mandatory to preserve digital cultural heritage in the form of web content. The Portuguese Web Archive project began in 2008. Since then, it has periodically collected live-web content to be preserved but also acquired historical web collections from third-parties previously published. However, storing information before it vanishes from the web is not enough to make web archives useful to societies. Thus, the Portuguese Web Archive developed and made freely available several software tools to enable access to web-archived collections. The Portuguese Web Archive provides a full-text search service to access 1 131 million files archived from the web since 1996 (`www.archive.pt`). It also provides access methods to enable research and development activities over web-archived data.

Keywords

Web archiving, digital preservation, Portuguese Web Archive

1. INTRODUCTION

The web is replacing printed media. Nowadays, we can find all kinds of publication genres transposed to online equivalents: electronic books, photo galleries, personal diary blogs, news articles, discussion forums or social networks. However, all this valuable information that describes our current days quickly disappears from the web [6]. In the same way that national archives or libraries have been preserving information published through printed media, it is now mandatory the creation of web archives to preserve the information published online.

Many web archives spread around the world collect and store web content [5]. However, enabling broad and efficient access to the web archived data is crucial to make web archives useful for societies. Live-web search engines are essential tools to enable access to current information. Web archives should complement live-web search engines by enabling access to past information published online.

The Portuguese Web Archive (PWA) project began in 2008 and aims to preserve information published on the web of main interest to the Portuguese community. Nonetheless, it also preserves content of international interest such as the sites from reputable worldwide organizations. In 2010, the PWA released the first version of a public search service that enables full-text search over its archived information. In March 2013, the PWA held 1 834 million web files collected from the web and integrated from historical collections provided by third-parties. Figure 1 presents the search interface that provides access to 62% (1 131 million) of the web files archived since 1996. This innovative service is publicly available at `archive.pt` and can serve a wide scope of user profiles and use cases. For example, web archive search can be useful to: journalists documenting articles, webmasters recovering lost versions of pages, historians studying digital documents about past events, lawyers obtaining evidence for legal cases, engineers consulting old documentation to fix legacy equipments or common web surfers recovering their broken-link favourites.

The software that supports the PWA was based on the Internet Archive Archive-access project tools [7], which are used by most web archives worldwide [5]. However, we observed that these tools did not fulfill our users requirements. Thus, we enhanced and adapted the Archive-access tools to support our service and made all the developed software available as a free, public open source project that can be reused and improved by other web archivists (available at `code.google.com/p/pwa-technologies/`). The PWA also provides tools and services to enable research over the archived data such as a distributed data processing platform or an OpenSearch API to facilitate the development of web applications that need to access web-archived data.

This demonstration will enable the attendees to discover the services provided by the PWA and discuss with the authors the details about how to develop and maintain a preservation service for web publications.

2. ACQUIRING HISTORICAL WEB COLLECTIONS

Since the first steps of the Internet in Portugal that individuals and organizations keep copies of published web content, most of the times for backup purposes. The PWA started collecting information for the web in January 2008 but we also acquired online content previously published to be preserved. We obtained historical content from web data sets gathered by research projects and personal collections that were gently supplied by their authors (see `arquivo.pt/supply` for details). The PWA preserves a total of 175 million web files (2.47 TB) supplied by third-party entities. The majority of these data was obtained from the Internet Archive that provided 124 million (1.9 TB) of data gathered from .PT between 1996 and 2007. Replicating these collections improved their chance of long-term preservation. Plus,

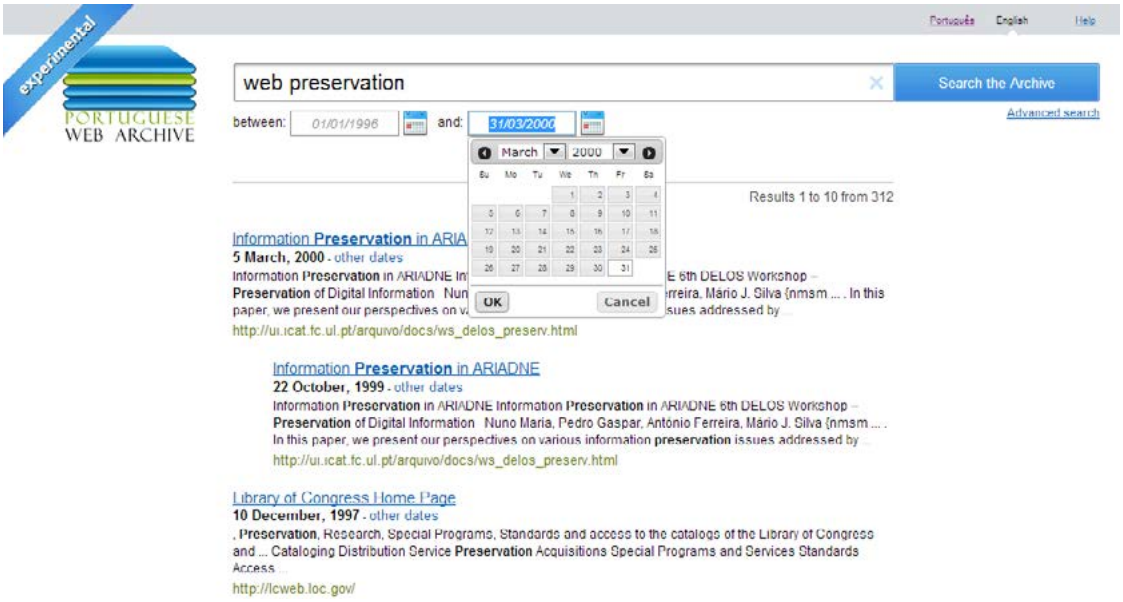


Figure 1: Result page for a full-text search over the Portuguese Web Archive (www.archive.pt).

they became full-text searchable through the PWA search service, while on the Internet Archive Wayback Machine the users have to know the exact address of the archived page that contained the desired information.

The acquired historical collections reached our web archive in heterogeneous formats, media support (e.g. CD-ROMs, backup tapes, original source code) and with scarce associated meta-data (e.g. missing original site URL or publication dates). Making this data searchable implied converting it to a uniform archive format so that it could be automatically processed indexed such as ARC or its successor standard WARC. We chose to use the ARC format instead of the official standard because it is the most widely supported by web archiving tools. The historical web collections acquired from the Internet Archive were delivered in the ARC format and were directly integrated in the PWA. The remaining acquired historical web collections were delivered to us in several distinct formats. Thus, we had to create specific integration modules to convert each one of these collections to the ARC format, which imposed a significant effort to integrate a relatively small amount of data. However, these collections provided valuable unique content. For instance, we obtained a version of the Library of Congress homepage dated from April 1996, while the oldest version existent on the Internet Archive is dated from December 1997.

A recurrent situation that we faced during the integration of the historical web collections was that acquired data were site backups made on local file systems with unknown or obsolete software. We observed that the majority of the acquired web collections created by organizations and individuals have been generated using software that was not designed with long-term preservation concerns, such as offline-browsers or through the Save feature of common web browsers. Offline browsers typically do not store meta-data related to each content saved locally, such as the original URL. As consequence, web archives cannot support URL search over these contents. If full-text is supported by a web archive, the contents could still be searchable but link-based algorithms could not be directly applied. Thus, these type of integrations required reverse engineering to model the archive file format and extract content meta-data. For instance, in 1995 a CD-ROM containing a snapshot of the Portuguese web was published as an attachment of a book. The web collection had the original URLs embedded as a reminder within each HTML page. However, non-HTML contents such as images did not have any URL associated. The extraction of the original URLs for these contents was automatized because each site was stored on a different directory and the original URLs were inferred by following relative links from pages. If the page with the URL `site.net/index.html` refered the image located in `./0.jpg`, then the original URL for the image was `site.net/0.jpg`.

As new web archiving initiatives arise they face the same challenge of having to integrate past content that is no longer available online and must be acquired from third-parties. Thus, we shared publicly the software developed to integrate our obtained historical web collections. The integration software was developed modularly so that it can partially applied and combined to address recurrent problems in independent collections. The converter of the CD-ROM collection to ARC format is available as an open source project at `code.google.com/p/roteiro2arc/`. HTTrack is a crawler used by web archives that stores content in a specific format [5]. The main purpose of HTTrack is to create site backups and the web collections generated with it contain most of the meta-data required to be successfully converted to the ARC format. The software to convert HTTrack crawls to ARC files is available at `code.google.com/p/httrack2arc/`.

3. PROVIDING ACCESS TO WEB-ARCHIVED CONTENT

Web archives contain rich and diverse information about international, regional or even personal events. However, web-archived information must be widely accessible to be useful. Most web archives rely on Archive-access tools to

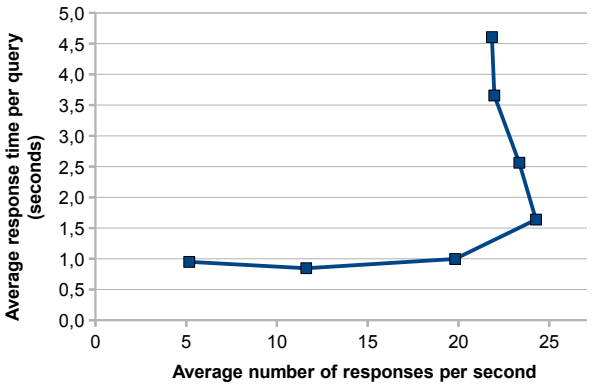


Figure 2: Experimental search performance results: relation between average response time and workload with load balancing across 7 machines.

provide access to their collections, in particular on the Wayback Machine for URL search and content visualization and NutchWAX for full-text search [5]. However, NutchWAX did not support the indexing of web collections that contained several versions of URLs harvested across time, which prevented its direct usage to index our acquired historical web collections that contained multi-version content. The performance of NutchWAX was also considered unsatisfactory by web archive stakeholders [5], missed support for internationalization of the user interface and did not include a query suggestions mechanism. We used the NutchWAX (v.0.11) and Wayback Machine (v.1.2.1) software as inception and enhanced it by addressing these drawbacks. The resultant software that supports the PWA search service was made publicly available as a free open-source project (`code.google.com/p/pwa-technologies/`). The introduction of a query suggestion mechanism had great impact on the perceived quality of our web archive because during usability testing we observed that users frequently mistyped queries and blamed the web archive for poor search results, often failing to spot their own mistypes [3]. The developed query suggestion mechanism was based on Hunspell, optimized for the Portuguese language and to the lexicon of our web collections [1] (`code.google.com/p/pwa-technologies/wiki/PwaSpellchecker`).

3.1 Search throughput capacity

Multiple, large and heterogeneous web collections must be searchable by queries in a few seconds. Users are used to the fast and high precision results of live-web search engines, such as Google, and expect the same behavior from web archive search systems.

We executed performance workload simulations to measure the throughput capacity and average response time of our web archive to search queries. These performance experiments were executed in a laboratory controlled environment to enable their reproducibility. The test collection was composed by 147 million web files gathered from 1996 to 2007. The experimental setup was composed by one load balancer that distributed the queries among 7 replicated search servers. The load balancing mechanism was implemented using the Linux Virtual Cluster software. Each search server supported queries over the full collection. The search servers shared the data storage device through a Storage Area Network that held the index. Each machine had 2 Xen Quad-core CPUs, 32 GB of memory and ran Linux. An increasing number of queries were submitted in parallel during a fixed interval of 5 minutes using several instances of the JMeter software and it was measured the time taken by PWA search system to respond to each query. The query set used to simulate the workload was composed by 300 000 queries obtained from a Portuguese web search engine [10] because a representative and structured web archive query log data set was not available at the time. Figure 2 presents the relation between workload and response time supported by the system. The obtained results show that up to an average workload of 20 responses per second, the system is able to maintain an average response time of approximately 1 second. However, when the workload reaches 25 responses per second, the average response time increases to 1.5 seconds and the system reaches its exhaustion point. From this point, we continued to increase the number of queries issued to the system but it was unable to respond to them. Thus, the system entered a thrashing state caused by overload and the average number of served responses per second decreased while the average response time increased.

Response time (s)	%full-text queries	%URL queries
$[0, 1[$	62.9%	71.7%
$[1, 2[$	14.9%	11.7%
$[2, 3[$	9.9%	6.5%
$[3, 4[$	4.5%	1.4%
$[4, 5[$	2.3%	2.2%
$[5, \infty[$	5.5%	6.5%

Table 1: Response time distribution derived from query log analysis (seconds).

The obtained results showed that our search software was able to support a throughput of 25 responses/second with an average response time of 2 seconds using load balancing across 7 machines. However, our search software may have to be installed on a smaller number of servers due to budget restrictions. Hence, we repeated the experiments and evaluated our search software without using any load balancing mechanism. We concentrated all software components and data structures on a single machine and the obtained results showed that the maximum supported throughput was 5 responses/second with an average response times of 2 seconds. Adding one replica and balancing the queries among these two machines, the response throughput increased to 10 responses/second with the same average response times of 2 seconds. We concluded that our search software is able to provide a satisfactory performance even when installed on limited hardware infrastructures.

The experimental setup previously described was deployed to production and we analyzed the logs of the queries issued by real users between May 2010 and July 2011 over a web collection of 187 million web files. Table 1 presents the response time distribution for full-text and URL queries. Around 87.7% and 89.9% of the full-text and URL queries, respectively, were responded in less than 3 seconds.

3.2 Search results relevance

Measuring the relevance of web archive search results usage requires test collections to obtain representative and reproducible results. However, existing test collections from

evaluation campaigns, such as the Text REtrieval Conference (TREC), do not address web archive requirements. For instance, their data sets are not composed by historical web collections gathered across time and the query sets are not focused on temporal queries that reflect the needs of web archive users. Therefore, we made a first effort to evaluate the relevance of our search results by performing a user click-through analysis derived from the query logs of the production environment of the PWA search service gathered from June to December 2010. The obtained results showed that 66% of the clicks were made on the first page of results, 23% of the clicks were made by the users on the first result presented by the system and 12% on the second result. These results are similar to those presented on web search engine studies [2]. Thus, they are a positive indicator of relevance. Another positive indicative is that only 2% of the URL search sessions did not receive any click by the users.

On the other hand, we obtained two negative quality indicators. The first one, is that 31% of the full-text search sessions did not receive any click by users. This abandonment rate suggests that users quit search before finding what they needed. The second negative indicative is that 85% of users identified by IP address did not revisit the web archive during the seven months period. One possible explanation for the non-revisit figure is that most users do not have a frequent need to search for historical web contents as they do for current information. Hence, the interval of time for users to revisit a web archive tends to be longer than for search engines. A longer time interval between revisits also reduces the probability of the same user revisiting the web archive using the same IP address.

3.3 Access tools for research

Web-archived information is a precious and abundant source of raw data for research. The PWA provides access tools and services to enable research over its archived data. The PWA has collaborated with researchers by providing data sets and access to its computing platform based on Hadoop. For instance, research has been performed to analyze the evolution of web characteristics [9], evaluate cross-lingual web classification algorithms [4] or to measure the accessibility of the web to people with disabilities [8]. To enable the automatic measurement of web page accessibility, we developed a software library that facilitates the development of distributed applications to process archived data. The PwaProcessor library interacts with the Hadoop framework for the distribution and the NutchWAX code for reading the content of ARC files (`code.google.com/p/pwa-technologies/wiki/PwaProcessor`).

Recently, the PWA published the first version of a test collection to support research on web archive information retrieval. It is composed by three parts: (1) a corpus representative of the documents' versions encountered in a real search environment; (2) a set of topics describing users' information needs; and (3) relevance judgments (a.k.a. qrels) indicating the degree of relevance of each document retrieved for each topic (`code.google.com/p/pwa-technologies/wiki/TestCollection`).

The PWA provides an OpenSearch access API that facilitates the development of web applications that need to access web-archived data (documentation at `code.google.com/p/pwa-technologies/wiki/OpenSearch`). This API has been frequently used by Computer Science students of the University of Lisbon to develop their academic projects.

4. CONCLUSIONS AND FUTURE WORK

Web archives already hold historical information spanning decades. However, making all these data widely accessible is still an open challenge. The Portuguese Web Archive collects information from the live web since 2008 and acquired historical web collections previously created by third-party entities.

The PWA search service enables full-text and URL search over 1 131 million files archived since 1996 (`archive.pt`). All the software developed to support this service was made publicly and freely available as an open source project (`code.google.com/p/pwa-technologies/`) so that it can be collaboratively enhanced and used as a baseline to develop more sophisticated accessible web archives in the future. The obtained experimental results showed that our search software is able to support a significant workload even when installed on limited hardware infrastructures and provide relevant search results. The PWA also provides services and data sets specially designed to support research and development activities.

5. REFERENCES

[1] M. Costa, J. Miranda, D. Cruz, and D. Gomes. Query suggestion for web archive search. In *Proceedings of the 10th International Conference on Preservation of Digital Objects (iPres 2013)*, September 2013.

[2] M. Costa and M. J. Silva. Characterizing search behavior in web archives. In *Proc. of the 1st International Temporal Web Analytics Workshop*, 2011.

[3] D. Cruz and D. Gomes. Adapting search user interfaces to web archives. In *Proceedings of the 10th International Conference on Preservation of Digital Objects (iPres 2013)*, September 2013.

[4] A. Garzó, B. Daróczy, T. Kiss, D. Siklósi, and A. Benczúr. Cross-lingual web spam classification. In *The 3rd Joint WICOW/AIRWeb Workshop on Web Quality in conjunction with WWW 2013*, WICOW 2013, May 2013.

[5] D. Gomes, J. Miranda, and M. Costa. A survey on web archiving initiatives. In *International Conference on Theory and Practice of Digital Libraries 2011*, Berlin, Germany, September 2011.

[6] D. Gomes and M. J. Silva. Modelling information persistence on the web. In *ICWE '06: Proceedings of the 6th international conference on Web engineering*, pages 193–200, New York, NY, USA, 2006. ACM Press.

[7] Internet Archive. Nutchwax - Home Page. `http://archive-access.sourceforge.net/`, March 2008.

[8] R. Lopes, D. Gomes, and L. Carriço. Web not for all: A large scale study of web accessibility. In *W4A: 7th ACM International Cross-Disciplinary Conference on Web Accessibility*, Raleigh, North Carolina, USA, April 2010.

[9] J. Miranda and D. Gomes. Trends in Web characteristics. In *7th Latin American Web Congress (LA-Web 2009)*, Merida, Mexico, November 2009.

[10] M. J. Silva. The Case for a Portuguese Web Search Engine. In P. Isaias, editor, *Proceedings of IADIS International Conference WWW/Internet 2003*, Algarve, Portugal, November 2003.

The SCAPE Planning and Watch suite

Supporting the preservation lifecycle in repositories

Michael Kraxner, Markus Plangg,
Kresimir Duretec, Christoph Becker
Vienna University of Technology
Vienna, Austria
{michael.kraxner, markus.plangg,
kresimir.duretec,
christoph.becker}@tuwien.ac.at

Luis Faria
KEEP Solutions
Braga, Portugal
lfaria@keep.pt

ABSTRACT

Increasingly, content owners are operating repositories with large, heterogeneous collections. The responsibility to provide access to these collections on the long term requires preservation processes such as planning, monitoring, and actual preservation operations such as migration and quality assurance, which have to be managed and integrated with the repositories. This article presents a suite of systems designed to support the preservation lifecycle in repositories. The SCAPE Planning and Watch suite provides the framework and toolset for controlling and monitoring scalable preservation operations. We present the main components for content profiling, preservation planning, and monitoring, and show how they can be combined to support scalable management of preservation over time.

Keywords

Digital Preservation, Preservation Planning, Preservation Watch, Content Profiling, Characterization, Scalability

1. MOTIVATION

The main focus of most digital repositories is to provide content access to its user community. To keep the content authentic and understandable to the user community on the long-term requires continuous monitoring, planning, and execution of corrective actions when needed to minimize risks and ensure continuous access. These processes need to be put together properly and integrated with repositories. The set of integrated digital preservation processes cover what we call the preservation lifecycle.

Scalability requires that automated tools support these processes, and a number of tools have emerged over the years to address parts of these processes. This ranges from aspects such as characterization, where tools such as JHove[1] and FITS[2] are commonly used, to preservation planning. The preservation planning framework and tool *Plato* provides a trustworthy method and support for decision making. A key challenge here is to scale the decision making support to enable responsible persons to manage large collections effectively and efficiently, and to integrate this support with repository systems. Additionally, continuous monitoring of the automated operations and the actual state of the repository and its content is needed to ensure that the repository's goals are met and risks and opportunities can be detected. This can only scale with automated tool support. Finally, a core aspect of this monitoring, and in effect the starting point for the preservation lifecycle, is an awareness of the the content itself and its various, potentially complex, and heterogeneous properties.

This demonstration presents an open, free, publicly available tool set that covers the crucial aspects of planning and monitoring outlined above and is integrated with a scalable environment for operations. We outline the key components and their conceptual interfaces, show how they address key issues of the preservation lifecycle, and demonstrate their interplay to address real-world preservation issues. All components are freely available on open licenses and can be accessed publicly on the web.

2. BACKGROUND

Digital preservation starts by knowing what content a repository has and what its key characteristics are. After digital objects have been characterized, this information can be aggregated and analysed and allows the content owner to get aware of the amount of content and the file format distribution. Moreover, this analysis process should support ways to drill down on important content characteristics to gain an in-depth understanding of preservation risks. Previous work has shown the value of format profiles across repositories [4], but was restricted to file formats only. Petrov showed that large-scale content profiling should not be restricted to the format label, but include more of the specific features that cause preservation issues, and that it is feasible to create and analyse large-scale in-depth profiles [7].

A preservation watch service has to cross-relate the results of this content characterization with institutional policies and external information about the technological, economic, social, and sometimes even political environment that provides the context of a repository. This allows for the identification of preservation risks and cost-reduction opportunities. Repository events such as ingest or download of content can also be useful for tracking producer and consumer trends and reveal preservation risks. The requirements on a preservation watch service have been described in [1].

These possible risks and opportunities should then be addressed by creating or revising a preservation plan. *Plato*, the preservation planning tool [2], guides the planner through

[1] `http://hul.harvard.edu/jhove/`
[2] `http://code.google.com/p/fits/`

Figure 1: Preservation lifecycle

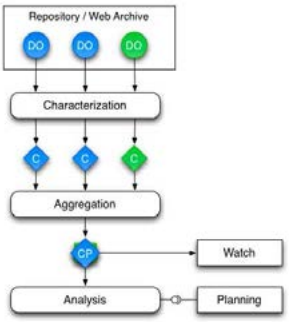


Figure 2: Content profiling

a well defined and approved workflow. After the organisational setting is described, decision criteria are defined, and representative samples as well as possible actions are selected, these actions are applied to the sample objects in controlled experiments, and the outcome is measured. Based on this, the planner decides which operation should be implemented.

This is a very solid, trustworthy process, but until recent improvements [3, 6] creating a plan used to be rather effort intensive, and sharing experience was difficult: Plans could be made public and exported to XML, but collaborative planning was only possible on public plans. Integrating such planning with the organisational context, the strategies and operations of an organisation was difficult, and monitoring changes was a manual process.

The SCAPE Planning and Watch suite has been developed to provide an open, scalable environment for preservation control and monitoring. It builds on the conceptual foundation of *Plato*, supports step-wise integration of systems through open interfaces, and enables organisations to practically apply Planning and Watch in a scalable, semi-automated way.

The next section will describe the key elements of the suite and their key features and relationships, while Section 4 gives an overview on upcoming improvements.

3. SCAPE PLANNING AND WATCH

First this section will give an overview on the preservation lifecycle and how all involved components work together. Then these components are described in more detail.

3.1 Preservation Lifecycle

Figure 1 illustrates the preservation lifecycle and how the the preservation components interact to support it.

The lifecycle starts with a repository containing content that is preserved for a designated community over time. Apart from this community, there is a wide variety of factors of interest to be monitored, including other repositories, technical solutions, format risks and other aspects [1]. The SCAPE Planning and Watch (PW) suite is designed so that in principle any repository can be connected. All interfaces are open, and a reference implementation for each API is being produced. This demonstration focuses on the current integration with the RODA repository[3].

[3] http://roda.scape.keep.pt/

In order to create a content profile, RODA includes a plugin that characterizes the files it is holding using FITS [4]. The results of characterization are fed into the scalable content profiling tool *c3po* [5]. *c3po* is using a scale-out NoSQL approach based on MongoDB[6] to provide highly scalable profiling of millions of objects, and creates a content profile that is exported to XML. This content profile is then linked with other aspects in the Watch component, *Scout* [5]. *Scout* can match these content profiles against organisational objectives and detect mismatches such as format profiles that pose specific risks, the presence of risk factors such as compression, or other conditions that are of interest and should lead to a mitigation.

A core enabler here is a semantic model for organisational policies and objectives that represents the drivers and constraints for preservation processes using an extensible ontology[7].

Upon discovering a condition that requires intervention, *Scout* notifies the responsible decision maker, who can use the visual analysis features of *c3po* to get an in-depth understanding of the issue at hand.

The result of preservation planning is a preservation plan, which contains an executable workflow. Upon completion and approval of the plan, it can directly be deployed to the configured endpoint of the repository. RODA contains a plugin to activate execution of such a plan on the original content set for which the content profile was created, thus closing the first circle.

Executing such operations on millions of objects will in the future be parallelized, for example on the SCAPE platform [8]. These operations will be monitored for compliance to the service level agreement, which will be done again using the monitoring component *Scout*.

3.2 Content profiling

c3po[8] enables in-depth analysis of the content of a repository. Figure 2 outlines the key steps of profiling. Content profiling covers aggregation and analysis of characterization data and distills it into a form suitable for planning and

[4] http://code.google.com/p/fits/
[5] http://www.ifs.tuwien.ac.at/imp/c3po
[6] http://www.mongodb.org/
[7] http://www.purl.org/DP/control-policy
[8] https://github.com/peshkira/c3po

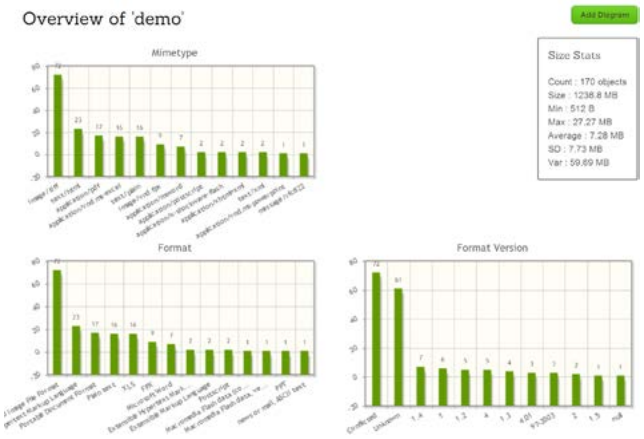


Figure 3: c3po showing a profile

watch. The aggregated data can then be used by other services like preservation watch, moreover it is visualised via a web front-end, as shown in Figure 3.

The distribution of digital object characteristics can be analysed and used to create sub-sets with certain properties, using interactive filtering on the charts themselves. Different algorithms are available to calculate representative samples.

This aggregated information can then be exported as a content profile and used for preservation planning and monitoring.

3.3 Preservation Watch

Scout[9], a preservation monitoring service, supports the scalable preservation planning process by implementing an automated service for collecting and analysing information on the preservation environment.

The information is collected by implementing different source adaptors. *Scout* has no restrictions on type of data that can be collected and it is planned to collect a variety of data from different sources like format- and tool registries, repositories, and policies. It already implements source adaptors for the PRONOM registry, content profiles from *c3po*, and policies. A repository adaptor, which is planned to monitor different events (ingest, access, migration) in a repository is currently being developed and with combination of content profiles from *c3po* will provide a complete overview of the current content in a repository and trends that are related to that content.

Once information is collected it is saved in a unified way to the knowledge base [5]. Built upon linked data principles the knowledge base supports reasoning on and analysis of the collected data. By providing different queries different information can be found. By now queries are used to provide a mechanism for automatic change detection. To do so a user simply deploys a trigger (a watch condition) which will be executed periodically. When the condition is met a notification is sent to the user. Upon receiving the notification the user can initiate additional actions like preservation

[9] `https://github.com/openplanets/scout`

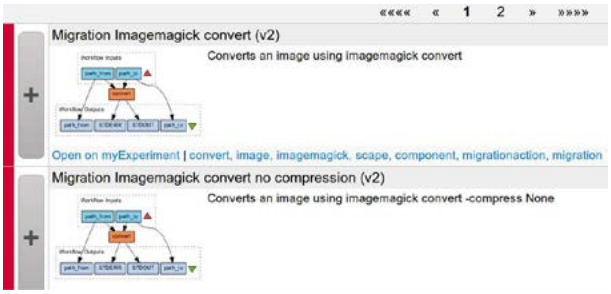


Figure 5: PLATO 4 querying myExperiment

planning.

Scout has a simple web interface which allows operations like management, adding new adaptors and triggers, and browsing the collected data. By operating over a longer period *Scout* is expected to have a valuable collection of historical data. In Figure 4 evolution of file formats through time is shown. The resulting graph is created by analysing approximately 1.5 million of files gathered in the period from December 2008 to December 2012 by the Internet Memory foundation[10].

3.4 Preservation Planning

Creation of a plan is supported by *Plato* 4[11], which is integrated with a number of aspects that provide substantial improvements in planning efficiency by reducing previously manual steps:

- The core understanding of the organisational drivers and constraints is provided by the semantic policy model which can be shared across members of the same organisation. This removes much of the contextual clarification that previously made starting a planning process difficult [3, 6]. When control policies have been defined, a preservation case can then be selected, and its information is applied to the plan. Decision criteria are derived from the objectives and mapped to the corresponding measures. Later quality assurance components will be looked up based on these measures, and the results can be applied automatically.
- Content profiles created by *c3po* are directly integrated with the planning workflow, so that both the analytical step of analysing content sets and the technical processes of characterization and selection of sample data for experimentation are fully automated.
- Figure 5 shows how discovery of applicable preservation actions can rely on preservation workflows shared and published using myExperiment [12], a social workflow sharing platform increasingly used by preservation practitioners. There are already a number of workflows for migration and quality assurance of image and audio files, some of them based on well-known tools like *FFmpeg*, others use new tools specifically developed for quality assurance, like jpylyzer[13].

[10] `http://internetmemory.org`
[11] `https://github.com/openplanets/plato`
[12] `http://myexperiment.org`
[13] `https://github.com/openplanets/jpylyzer`

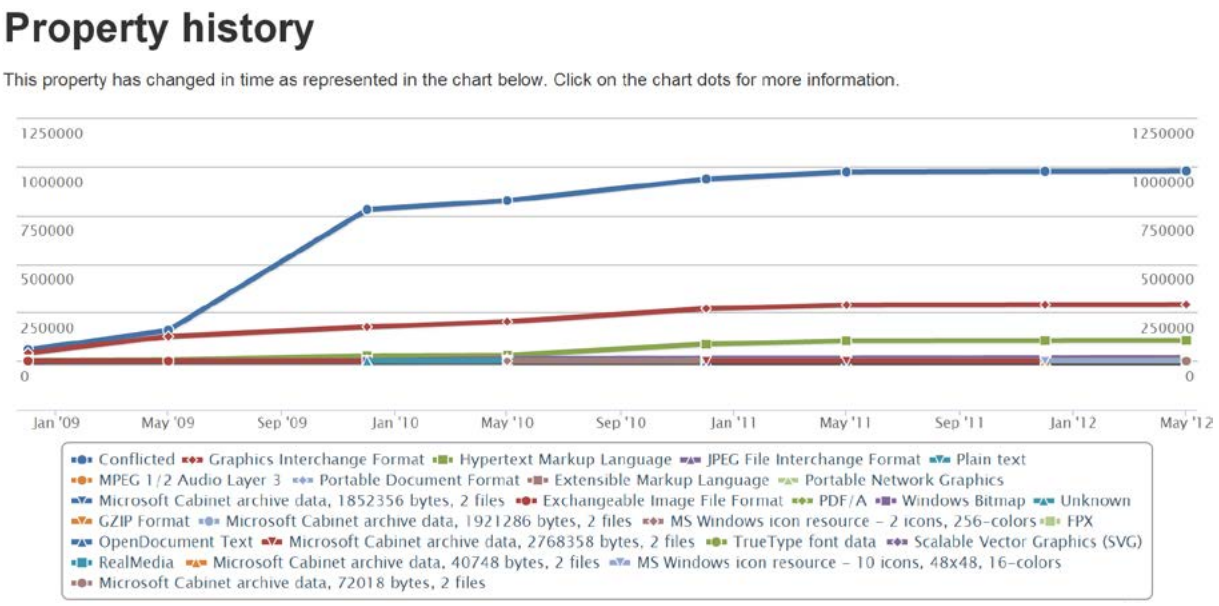


Figure 4: Property values over time in Scout

- Experiment execution is highly eased by integrating the Taverna workflow engine [14] which is used to run the candidate preservation actions. The content profiles reference the permanent identifiers of the original objects, so that the byte-streams of sample objects can be directly used to test applicable actions on examples of the actual dataset that should be preserved.
- After the outcomes of the preservation actions have been measured, the results have been evaluated, and the best alternative has been identified, a Preservation Action Plan can be created based on this workflow and the content profile, and once the plan is approved, it can be deployed to the configured repository endpoint, where it can be executed.

This is a major step upwards from previous iterations, where policies were implicit, content profiles manual, requirements specification effort-intensive, action discovery limited and plan deployment manual.

4. SUMMARY AND OUTLOOK

This demonstration presents a suite of systems that enable scalable monitoring and control of preservation in real-world environments. While each tool can be (and is) used independently, they are designed to be highly interoperable, so that the compound value contribution is larger than the sum of its parts. Using the SCAPE Planning and Watch tool suite, we can manage and streamline the continuous execution of digital preservation processes (the preservation lifecycle) on a systematic and semi-automatic way, mitigating some of the problems of large-scale digital preservation in an effective way.

As a next step the introduced APIs will be published, so they can be adopted by parties outside of SCAPE. Annotated SCAPE components will improve lookup, and ease composition, which enables to improve automation of quality assurance as well as generation of Preservation Action Plans.

[14] `http://www.taverna.org`

Acknowledgements

Part of this work was supported by the European Union in the 7th Framework Program, IST, through the SCAPE project, Contract 270137.

5. REFERENCES

[1] C. Becker, K. Duretec, P. Petrov, L. Faria, M. Ferreira, and J. C. Ramalho. Preservation watch: What to monitor and how. In *Proc. IPRES*, 2012.

[2] C. Becker, H. Kulovits, A. Rauber, and H. Hofman. Plato: A service oriented decision support system for preservation planning. In *Proc. JCDL*, 2008.

[3] C. Becker and A. Rauber. Preservation decisions: Terms and Conditions Apply. Challenges, Misperceptions and Lessons Learned in Preservation Planning. In *Proc. JCDL 2011*, June 2011.

[4] T. Brody, L. Carr, J. Hey, A. Brown, and S. Hitchcock. Pronom-roar: Adding format profiles to a repository registry to inform preservation services. *IJDC*, 2(2), November 2007.

[5] L. Faria, C. Becker, P. Petrov, K. Duretec, M. Ferreira, and J. Ramalho. Design and architecture of a novel preservation watch system. In *Proc. ICADL*, 2012.

[6] H. Kulovits, C. Becker, and B. Andersen. Scalable preservation decisions: A controlled case study. *Archiving 2013*, 2013.

[7] P. Petrov and C. Becker. Large-scale content profiling for preservation analysis. In *Proc. IPRES*, 2012.

[8] R. Schmidt. An architectural overview of the SCAPE preservation platform. In *Proc. IPRES*, 2012.

Demonstration of the BitCurator Environment

Christopher A. Lee
School of Information and Library Science
University of North Carolina
216 Lenoir Drive, CB #3360
1-(919)-966-3598
callee@ils.unc.edu

ABSTRACT
This demonstration will illustrate the design and functionality of the BitCurator environment, which is a set of open-source software that allows collecting institutions to apply digital forensics methods to digital materials. These methods help to advance a variety of digital preservation goals.

Categories and Subject Descriptors
H.3.7 [**Information Storage and Retrieval**]: Digital Libraries – *collection, dissemination, systems issues.*

General Terms
Provenance, Data Triage, Digital Forensics.

Keywords
Forensics, preservation, DFXML, metadata, privacy, collections, acquisition

1. BITCURATOR PROJECT
The BitCurator Project, a collaborative effort led by the School of Information and Library Science at the University of North Carolina at Chapel Hill and Maryland Institute for Technology in the Humanities at the University of Maryland, builds on previous work by addressing two fundamental needs and opportunities for collecting institutions: (1) integrating digital forensics tools and methods into the workflows and collection management environments of libraries, archives and museums and (2) supporting properly mediated public access to forensically acquired data [4].

2. BITCURATOR ENVIRONMENT
We are developing and disseminating a suite of open source tools. These tools are currently being developed and tested in a Linux environment; the software on which they depend can readily be compiled for Windows environments (and in most cases are currently distributed as both source code and Windows binaries). We intend the majority of the development for BitCurator to support cross-platform use of the software. We are freely disseminating the software under an open source (GPL, Version 3) license. BitCurator provides users with two primary paths to integrate digital forensics tools and techniques into archival and library workflows.

First, the BitCurator software can be run as a ready-to-run Linux environment that can be used either as a virtual machine (VM) or installed as a host operating system. This environment is customized to provide users with graphic user interface (GUI)-based scripts that provide simplified access to common functions associated with handling media, including facilities to prevent inadvertent write-enabled mounting (software write-blocking).

Second, the BitCurator software can be run as a set of individual software tools, packages, support scripts, and documentation to reproduce full or partial functionality of the ready-to-run BitCurator environment. These include a software metapackage (.deb) file that replicates the software dependency tree on which software sources built for BitCurator rely; a set of software sources and supporting environmental scripts developed by the BitCurator team and made publicly available at via our GitHub repository (links at http://wiki.bitcurator.net); and all other third-party open source digital forensics software included in the BitCurator environment.

3. DEMONSTRATED TOOLS AND FEATURES
Tools that BitCurator is incorporating include Guymager, a program for capturing disk images; bulk extractor, for extracting features of interest from disk images (including private and individually identifying information); fiwalk, for generating Digital Forensics XML (DFXML) output describing filesystem hierarchies contained on disk images; The Sleuth Kit (TSK), for viewing, identifying and extraction information from disk images; Nautilus scripts to automate the actions of command-line forensics utilities through the Ubuntu desktop browser; and sdhash, a fuzzing hashing application that can find partial matches between similar files. For further information about several of these tools, see [1,2,3,5].

This demonstration will illustrate BitCurator support for mounting media as read-only, creating disk images, using Nautilus scripts to perform batch activities, generation of Digital Forensics XML (DFXML), generation of customized reports, and identification of sensitive data within data.

4. ACKNOWLEDGMENTS
BitCurator development has been supported by the Andrew W. Mellon Foundation. Members of the BitCurator team are Alexandra Chassanoff, Matthew Kirschenbaum, Christopher (Cal) Lee, Sunitha Misra, Porter Olsen, and Kam Woods. Members of two advisory boards have made valuable contributions: the Development Advisory Group (DAG) and Professional Experts Panel (PEP).

5. REFERENCES

[1] Cohen, M., Garfinkel, S., and Schatz, B. 2009. Extending the Advanced Forensic Format to Accommodate Multiple Data Sources, Logical Evidence, Arbitrary Information and Forensic Workflow. *Digital Investigation* 6 (2009), S57-S68.

[2] Garfinkel, S. Digital Forensics XML and the DFXML Toolset. *Digital Investigation* 8 (2012), 161-174.

[3] Garfinkel, S.L. Providing Cryptographic Security and Evidentiary Chain-of-Custody with the Advanced Forensic Format, Library, and Tools. *International Journal of Digital Crime and Forensics* 1, 1 (2009), 1-28;

[4] Lee, C.A., Kirschenbaum, M.G., Chassanoff, A., Olsen, P., and Woods, K. BitCurator: Tools and Techniques for Digital Forensics in Collecting Institutions. *D-Lib Magazine* 18, 5/6 (May/June 2012).

Roussev, V. An Evaluation of Forensic Similarity Hashes. *Digital Investigation* 8 (2011), S34-S41.

Using data archiving tools to preserve archival records in business systems – a case study

Neal Fitzgerald
GLAMATEK
Brisbane, Queensland, Australia
neal.fitzgerald@gmail.com

ABSTRACT

The preservation of archival records from government business systems is a pressing concern for archival institutions worldwide. Most business systems developed over the last 20 years do not have in-built recordkeeping functionality. Archivists and records managers face the task of identifying, extracting and preserving archival records from these systems. I use a public authority collection management system as a case study to explore how currently available archiving tools might form part of a practical method to identify, extract and prepare digital archival records for ingestion into a digital preservation archive.

Keywords

Digital preservation, SQL Server, databases, business systems.

1. BUSINESS SYSTEM ARCHITECTURE

The most common business system architecture is an application layer built on top of a commercial relational database layer. The database layer holds the system data in related tables and usually some code in *stored procedures* and *triggers* that perform database access and update functions. The application layer is made up of code modules that contain the business rules, manage workflows and generate user interface screens. It presents the data in different ways, summarizes it and produces reports. It turns the data into information. You need the application layer to make full sense of the database layer [4].

Database management systems use Structured Query Language (SQL) to access and manipulate data stored in rows and columns of related tables. Databases are designed to minimize redundant or repeated data. It is common to use codes to link to common values stored in separate lookup tables. Repeating fields and transactions are also decomposed into separate tables. *Views* are database objects that store a single SQL query. They are often used to bring back together decomposed elements of common system entities. They act like virtual tables that can be referenced in database and application code.

2. RECORDS IN BUSINESS SYSTEMS

Generally business systems are not recordkeeping systems. Data is fluid and changing. Historical data is often overwritten to reduce storage costs and keep the system running efficiently so it is hard to reconstruct the system state at any point in time. Historical data that is maintained is often not tamper proof [4]. Historical reports, summaries, snapshots or extracts created by the business system are often held outside the database. These could be printed on paper, or held in a file system, data warehouse or eDRMS (electronic document & records management system).

In the IT world a record is a row in a database table. To archivists records are information created, received and maintained as evidence and information by an organisation or person, in pursuance of legal obligations or in the transaction of business [ISO 15489]. I use the word in this sense throughout this paper.

In Queensland, a *public* record is any form of recorded information, either received or created by a public authority, which provides evidence of the business or affairs of that public authority [5]. Public records may need to be retained permanently or expire after a fixed period (the *retention period*).

The International Council on Archives has developed *Guidelines and Functional Requirements for Records in Business Systems* to make design suggestions for new systems and as a way to review recordkeeping functionality in existing systems. The guideline states that to identify records as evidence we need to:

1. Determine the broad business functions and specific activities and transactions carried out by the business system.
2. For each function, activity and transaction or business process managed by the system, consider what evidence is required to be retained by the organisation.
3. For each requirement for evidence, identify the content or data that make up the evidence.

Records might consist of a number of inter-related data elements connected across one or more tables [3].

3. THE PROBLEM

A recent Queensland Government ICT audit identified a number of business systems for decommissioning. Government wants to reduce the cost of software licensing, hardware maintenance and specialist skills. Systems containing data with ongoing business use will be kept running or virtualised until the data is archived, migrated or no longer required. Systems with data that is no longer currently used are to be switched off as soon as possible. Agencies need to identify the records in the business systems that have not expired and are still within their retention period. If all the records in the system have expired, then the systems can be switched off without further action. For systems with unexpired or permanent records, agencies need processes and tools to extract these records in a format that can be preserved and rendered for long term access.

4. TOOLS & APPROACHES

A common approach to preserving business system records is to export the contents of all of the tables in the database in an open XML format. If records must be deleted after their retention period expires (for example some criminal records), unwanted data must be deleted from the database before archiving or from the archive package after archiving. Tools like RODADB and SIARD that use this approach have functions to load the XML archive to a different SQL database platform to allow ongoing

access over time, but this requires knowledge of the database structure and SQL query skills.

Commercial database archiving software tools like HPAIO and CHRONOS [1] are primarily designed to purge data from large transactional databases to reduce storage costs and improve performance. They use a similar export-all-tables approach for retiring business systems, but they also have functionality to assemble 'data objects' (and so archival records) from their constituent columns and tables and extract these in XML format. If we preserve these archival records, users do not need knowledge of the database structure or SQL skills to access them.

Database warehousing software is used by some agencies for enterprise level reporting, trend analysis and data mining across data assets from different business systems. It may be possible to leverage this software to extract and preserve records [4].

In a recent blog post [6] State Records of NSW discuss a number of methods of preserving and presenting the information in business systems to suit different classes of users. A searchable collection of pdfs might suit family researchers. An open source SQL database might suit agency IT staff to re-create reports from an archived system. RDF linked open data might suit a researcher wanting to create visualizations of the data.

In business systems applications, records as evidence of business transactions will often be presented to the user on a single screen or a set of related screens. At the National Archives of Sweden the 'preservation object' in a business system is documented with a screen shot, a mapping of the screen fields to database table columns, and the corresponding SQL query [2]. If the assembled archived records in XML format can be rendered in a form similar to the original business application screen including the field values, screen labels, field ordering and grouping, this would provide a human friendly way to present the records for access and also provide a visual check of the record accuracy.

5. RECORD PRESERVATION MODEL

I propose a four stage preservation model illustrated in Figure 1.

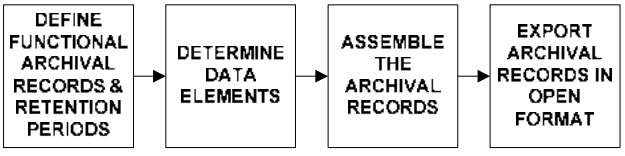


Figure 1. Preserving archival records in business systems.

1. Identify the archival records by analyzing the functions of the system that assist decision making or document business processes or actions and determine their retention periods.

2. Determine the data elements that document the system transactions performed to fulfill these functions by reviewing system documentation, application code, database structures and user interface screens and interviewing knowledgeable staff.

3. Assemble the data elements to produce a consolidated representation of the archival record.

4. Extract and export the consolidated records whose retention periods have not expired in an open format.

6. THE CASE STUDY

This case study uses an agency collection management system. It has a staff module for maintenance of the collection catalogue and a web based search interface for public access to the collection.

6.1 Defining the functional archival records

The agency has a retention and disposal schedule, which specifies the agency archival records held in the collection management system at a high level and their retention periods. One set of records is the digitized image register which I use as an example. The retention and disposal schedule specifies that the metadata describing the images needs to be preserved until superseded.

6.2 Determining the data elements

6.2.1 Reviewing the application screens

My starting point for determining the data elements making up the image register records was the corresponding enquiry screen in the application staff module. This screen showed descriptive metadata, staff who created and approved the image, and virtual exhibitions in which it appears.

6.3 Assembling the archival records

A 19.7Gb backup of the Microsoft SQL Server database layer was restored into a Microsoft SQL Server 2008 R2 installation on a Dell Optiplex 990 PC with Intel core i5-2400 @3.1GHz with 8Gb of RAM, running the Windows 7 64-bit operating system.

I used the SQL Server Management Studio tool to construct an SQL query to model the image register archival record. As I proceeded I compared the query results against values displayed for the example images on the image register enquiry screen.

6.3.1 Reviewing the database

The database has over 300 tables. Table and column names are generally descriptive. There are no entity relationship (ER) diagrams and no declared foreign keys, so I looked at database and application code for clues to the table relationships.

Reviewing the table names, I found a candidate main table for the image register record query with column names matching the screen labels and two large object binary columns containing jpeg images. These images are access copies and do not need to be preserved in the record. Preservation TIFF versions are kept elsewhere on the file system.

A simple SQL query on this table returned some values matching those on the screen and some codes. I found a stored procedure that displays image metadata for the public web interface. It contained an SQL query that showed joins to some of the lookup tables. I added these tables to my record query and the results matched the screen except for the staff and exhibitions data.

6.3.2 Reviewing the application code

I found the application module that produced the image register screen. It contained somewhat cryptic and fragmented Visual Basic code that constructed an SQL query. It used the same joins to the lookup tables found in the database layer stored procedure. The code also showed the joins to the staff lookup and exhibitions tables. I added these joins to the SQL query and it returned all data as displayed on the image register screen.

6.3.3 Creating the archival record table

I created a *view* object to document my query and used the query to make a new database table with all the elements of the record.

Tables in relational databases have 1-to-1, 1-to-many or many-to-many relationships. Many-to-many relationships are usually decomposed into two 1-to-many relationships via a chaining table. The relationship between image register table and code lookup

tables are one to one. A single image code field corresponds to one value in the lookup table. The relationship between the image register and the exhibition tables is many-to-many. We can represent this by two 1-to-many relationships. An image may be in more than one exhibition. An exhibition has a number of images. The image screen shows the image details and lists all the exhibitions in which it appears in a scrolling window and the exhibition screen shows the exhibition details and lists its images.

When tables with a 1-to-many relationship are joined in an SQL query, the result is a *Cartesian product* of the two tables. In our example, if an image occurs in a number of exhibitions, there will be a row for each exhibition and the image data will be repeated with each row of the result. Because images rarely appear in multiple exhibitions, the output from the record query in this case is not significantly larger than the size of the original tables.

In other cases there may be a large number of child rows for a large number of parent rows. An assembled record table could be orders of magnitude larger and add significantly to the archive size. I used the SQL Server XML functions with my query to create XML with a single copy of the image data for each image and nested exhibition elements. With the full dataset the query failed with a memory error, common in my experience on a number of platforms. An alternative is to use XSLT style sheets after archiving to create hierarchically structured XML.

6.4 Exporting archival records

I was able to download trial versions of SIARD, RODADB and HPAIO. CHRONOS staff demonstrated their software by WebEx. The case study database had to be prepared to allow the tools to connect by enabling TCP/IP, opening ports, starting services and enabling SQL Server authentication.

6.4.1 HP Application Information Optimizer[1]

HPAIO was formerly known as HPDBA. HPAIO version 7.02 is complex to install and operate. It archives data as XML or CSV documents. It can export these documents to HP's TRIM eDRMS. It is quite complex to install and operate and has a lot of functionality not required in our decommissioning use case.

The HPAIO Designer tool lets you create *models* of *business objects* using a visual drag and drop interface to join the database tables. You select a driving table and add lookup, chaining or transactional tables, essentially creating the equivalent of an SQL query. You can select a subset of columns in each table, rename table columns and add selection conditions. A pdf document can be produced listing the components that make up the model.

I used this tool to model the image register archival records using the tables and joins discovered during the previous analysis. HPAIO produced a nested hierarchical XML. The extract failed for the full record set with a memory error. HPAIO embeds binary image files within the XML, which would hinder monitoring these embedded objects for format obsolescence over time.

6.4.2 CHRONOS

CHRONOS[2] database XML export format is simple and compact. It shows table structures, the queries in view objects, and the code in stored procedures and triggers. Each table is stored in a compressed CSV zip file with checksums stored for each row. It can export a *view* object as if it was a real table, assembling the data and creating a CSV file. If we create a *view* from our archival record this gives a simple record extraction method. Exporting the records as tables with the more compact compressed CSV format rather than XML reduces the impact of the Cartesian product problem described in 6.3.3 above. CHRONOS has user access control, functionality to support WORM storage devices and SHA-512 encryption to increase security and prevent data tampering. We hope to have a pilot installation to test soon.

6.4.3 SIARD / SIARDK

The Swiss Federal Archives publishes the SIARD[3] (Software Independent Archiving of Relational Databases) open standard database archiving XML format. They provide free tools to export SQL databases to SIARD XML and tools to import the SIARD XML into various database management systems for access.

The SIARD archive package is a ZIP64 packaged hierarchy of folders and files. The *header* folder contains the SIARD schema, a *metadata.xml* file (describing the database tables, views and stored procedures), and an XSL style sheet for viewing the package contents in a web browser. In the *content* folder there is a folder for each table with the data in *table.xml* and a schema describing the table structure in *table.xsd*. ZIP64 allows for large packages whereas the ZIP format is restricted to 4Gb.

I installed SIARD version 1.50. The tool is java based and easy to install and use. The database was converted to an 18.7 Gb SIARD ZIP64 package in about 4 hours. There is a graphical user interface for inputting the connection parameters and initiating export and import tasks. The graphical user interface allows browsing of SIARD archive packages. It also allows extra descriptive metadata to be added to the package.

SIARD stores large object binary elements such as large text blocks, images or video as separate files in the archive package referenced from the XML. SIARD does store the contents of database views, but did not store any code from stored procedures, only their name and parameter declarations. Potentially valuable information about database structure and system function is lost.

6.4.4 RODADB

The RODADB command line tool has been derived from the database ingestion and deployment components of the RODA digital preservation repository[4]. It exports the database table definitions and table content for all the tables in the database in DBML XML format. The software also allows DBML to be converted to a set of SQL statements that can be used to re-create the tables in another database platform for access. I downloaded RODADB version 1.1.1. It is java based application and easy to install. The command line syntax is straight forward. RODADB also stores the XML and large object binary images separately. The output is a folder containing a single XML file with pointers to the image files in the same folder. The case study database export produced a 21Gb XML file and about 90,000 binary image files (12Gb) in about 3.5 hours.

---

[1] http://www8.hp.com/au/en/software-solutions/software.html?compURI=1175612#.UXTEHyJMj4-

[2] http://www.csp-sw.de/en/inhalt.php?kategorie=c271_CHRONOS

[3] http://www.bar.admin.ch/dienstleistungen/00823/00825/index.html?lang=en

[4] http://www.keep.pt/produtos/roda/?lang=en

The DBML contains only table and key definitions and table data. Views and stored procedures are ignored. The lack of a graphical user interface might discourage some users. The DBML XML file is significantly larger than the corresponding SIARD XML files, but more directly readable, because it spells out the table column names for each row.

7. RECOMMENDATIONS

We should preserve the database content in a format that can be rendered in a number of ways for presentation. This way we can satisfy the needs of different classes of users from members of the public who want a Google like search and easy presentation through a web browser, to agencies who want to manipulate archived data in a similar way to when the business system was still functioning, to future researchers who want datasets to mine, visualize and mash up with other datasets [6].

The preservation effort expended on any business system will depend on many factors including legal risk, information value, available skills, available resources and historical importance.

For business systems being decommissioned:

- Use archiving tools that produce XML
- Create a full XML export of database tables
- Preserve as much contextual documentation as possible
- Identify and assemble archival records before archiving
- Document the archiving process
- Optionally preserve the original database layer backup file

XML is an open, text based format that does not require specialist software to be rendered, so has a good chance of remaining accessible in the long term. It can easily be transformed using XSL style sheets to allow human friendly display formats for access, and the creation of open data sets for data mining and visualization. There are tools to load SIARD XML and RODA DBML to a number of SQL database platforms for access by those with SQL query skills.

A full database export will decrease the risk that important information is lost unintentionally. In some cases some data may be mandated for deletion and some duplicate or ephemeral data may need to be deleted before archiving to save storage costs.

Contextual documentation might include user manuals, screen shots, application and stored procedure code, database ER diagrams, application architecture or UML diagrams, retention and disposal schedules, records of interview with IT staff and expert users. System reports and summaries produced by the business system in the past may be useful artefacts.

Identifying and assembling archival records from their constituent columns and tables in the database before archiving will aid future accessibility. The process is an opportunity to gather together, synthesize and crystallize information from contextual documentation, application code, database layer code and expert knowledge. The dataset modeling tools in HPAIO can be used to achieve this. If using SIARD or RODADB, an SQL query can be used to assemble the record elements into a new database table before archiving. The developed query can also be documented and stored as a *view* object in the database layer. Some systems will have better documentation than the case study with database ER diagrams, declared foreign keys and views or stored procedures that correspond to the archival records. The process may be easy and straight forward.

If storage permits the original database backup file could also be preserved, so the archiving process itself can be verified while the database software and operating environments are still available.

Documenting the archiving process in detail will give future researchers confidence that the archived records are a true representation of the original records in the business system.

If budgets are tight and a simple solution is needed, I would recommend SIARD as the initial tool of choice. If possible, spend time to identify and assemble the elements of the archival records in new tables before archiving as SIARD XML.

Of the commercial products reviewed, CHRONOS looks very promising. It exports more elements of the database in more compact and open way. If an agency implementing the system has a requirement to archive records and other data from currently functioning business systems, the investment in licenses and learning how to use the product would be worthwhile.

8. FUTURE WORK

I propose to further test and refine this work using business systems implemented on other database platforms. I will further experiment with the tools examined here and with other open source and commercial data archiving and data warehousing tools including CHRONOS. I will explore using XSLT style sheets to transform the XML exported from the case study database using the tested tools. Use cases will include rendering human friendly access versions of the records, removing unwanted data, creating hierarchical XML representations of assembled archival records that have repeating data, preparing datasets for migration to a new collection management system, creating Submission Information Packages to ingest into a digital preservation archive.

9. REFERENCES

[1] Brandl, S., Keller-Marxer, P. 2007. *Long-term Archiving of Relational Databases with Chronos.* Proceedings of the PresDB'07 workshop, Edinburgh.

[2] Geber, M. 2012. *Database Archiving in Sweden*, Presentation at 'A Practical Approach to Database Archiving' workshop, Copenhagen.

[3] International Council on Archives 2008. *Principles and Functional Requirements for Records in Electronic Office Environments – Module 3: Guidelines and Functional Requirements for Records in Business Systems*.

[4] O'Kane, T., Somerville, C. *Data in Databases – it's not what you think.* Presentation at Future Perfect 2012, Wellington.

[5] Queensland State Archives 2013. *What is a Public Record?* http://www.archives.qld.gov.au/Recordkeeping/GRKDownloads/Documents/what_is_public_record_200409.pdf.

[6] State Records of NSW Future Proof blog 16 June 2013. *Migrating Business Systems to the Digital Archives* http://futureproof.records.nsw.gov.au/migrating-business-systems-to-the-digital-archives-a-post-from-the-digital-archives-team/

PERICLES - Promoting and Enhancing Reuse of Information throughout the Content Lifecycle taking account of Evolving Semantics

Mark Hedges, Simon Waddington[1]
King's College London
{mark.hedges, simon.waddington}@kcl.ac.uk

Jens Ludwig, Philipp Wieder[5]
Georg-August-Universität Göttingen
ludwig@sub.uni-goettingen.de, philipp.wieder@gwdg.de

Rob Baxter[9]
Edinburgh Parallel Computing Centre
r.baxter@epcc.ed.ac.uk

Sándor Darányi, Elena Maceviciute, Tom Wilson[2]
University of Borås
{sandor.daranyi, elena.maceviciute}@hb.se, wilsontd@gmail.com

Paul Watry, Adil Hasan, Fabio Corubolo[6]
University of Liverpool
P.B.Watry@liverpool.ac.uk, {adilhasan2, corubolo}@gmail.com

Pip Laurenson[10]
Tate
pip.laurenson@tate.org.uk

Yiannis Kompatsiaris, Stamatia Dasiopoulou[3]
Centre for Research and Technology Hellas
{ikom, dasiop }@iti.gr

Rani Pinchuk[7]
Space Applications Services NV
rani.pinchuk@spaceapplications.com

Christian Muller[11]
B.USOC
christian.muller@busoc.be

Odysseas Spyroglou[4]
Dotsoft
ospyroglou@dotsoft.gr

Jean-Pierre Chanod, Jean-Yves Vion-Dury[8]
Xerox
{jean-pierre.chanod, jean-yves.vion-dury}@xrce.xerox.com

ABSTRACT
This poster paper describes the objectives, approach and use cases of the EC FP7 Integrated Project PERICLES. The project began on 1st February 2013 and runs for four years. The aim is to research and prototype solutions for digital preservation in continually evolving environments including changes in context, semantics and practices. The project addresses use cases focusing on digital art, media and science.

Categories and Subject Descriptors
Information Systems [**Information Systems Applications**]: Digital Libraries and Archives

General Terms
Theory.

Keywords
Preservation models, lifecycle, data analytics, semantics, policies.

1. INTRODUCTION
This poster paper describes the objectives, approach, use cases and proposed deliverables of the EC FP7 Integrated Project

PERICLES: http://www-pericles-project.eu. The PERICLES project was funded through the FP7 ICT Call 9 Digital Preservation. The project involves partners of a range of complementary types, including six academic partners, one multinational corporation, two SMEs and two non-academic public sector organisations.

2. PROBLEM DESCRIPTION
As digital content and its related metadata are generated and used across different phases of the information lifecycle, and in a continually evolving environment, the concept of a fixed and stable 'final' version that needs to be preserved becomes less appropriate. As well as dealing with technological change and obsolescence, long-term sustainability requires us to address changes in context, such as changes in semantics - for example, the 'semantic drift' that arises from changes in language and meaning - or disciplinary and societal changes that affect the practices, attitudes and interests of the 'stakeholders', whether these be curators, artists, scientists, or indeed a broader public, such as visitors to exhibitions.

Such a changing environment necessitates a corresponding evolution of the strategies and approaches for preservation if stakeholder communities are to be able to continue to use and interpret content appropriately. A key issue is the provision of sufficient contextual information to enable both lifecycle management and preservation on the one hand, and re-use or re-

interpretation of content on the other, as well as the facility to model and describe preservation processes, policies and infrastructures as they themselves evolve. Capturing and maintaining this information throughout the lifecycle, together with the complex relationships between the components of the preservation ecosystem as a whole, is key to an approach based on 'preservation by design', through models that capture intents and interpretative contexts associated with digital content, and enable content to remain relevant to new communities of users.

The project will address these preservation challenges in relation to digital content from two quite different domains: on the one hand, digital artworks, such as interactive software-based installations, and other digital media from Tate's collections and archives; on the other hand, experimental scientific data originating from the European Space Agency and International Space Station.

3. AIMS AND OBJECTIVES

The project has three main objectives.

Objective 1: To enable trusted access to digital content that is complex, heterogeneous, highly-interconnected, and subject to change, and to facilitate continued understanding and reuse of those objects across all phases of the lifecycle. This will be achieved by:

- Developing a model based on a linked data paradigm for describing the resources in preservation environment - including content, metadata, processes, users, and policies - and for managing their dependencies and consistency as the environments evolve.
- Adapting and extending preservation and lifecycle models to address the evolution of digital ecosystems and their dependencies, and developing an associated framework and tools.
- Developing a range of analytical methods for identifying and capturing preservation-related information - semantics, users, interpretative contexts - from digital content and its environment.

Objective 2: To evaluate our approaches, processes and tools against requirements and user communities in different application domains. This will be achieved by:

- Developing case studies to evaluate the approaches taken by PERICLES against the requirements of the user communities targeted by the project.
- Assessing the potential for deploying project outputs in operational environments.

Objective 3: To facilitate sustainability and exploitation of project outputs by disseminating the knowledge created by the project, and in particular by:

- Building communities of practice around a number of topics addressed by PERICLES.
- Engaging with standardisation activities regarding contribution to relevant standards.
- Engaging with commercial organisations to facilitate the take-up of project outputs by industry.

4. APPROACH AND METHODOLOGY

4.1 Case studies and evaluation

The research carried out by PERICLES will be driven by and evaluated against two distinct groups of case studies focused around different application domains and communities in media and science. While on the surface very different, these two areas have in common an environment that evolves continually, not only in terms of the technologies used, but also as regards meaning, and the practices, attitudes and interests of stakeholders, whether these be curators, artists, scientists, engineers, or indeed a broader public, such as visitors to exhibitions. By addressing the preservation challenges raised by digital material from two quite different domains we aim to ensure that our results are of broad applicability.

Rather than a single system, PERICLES will produce a variety of components (models, tools, policies, architectural approaches etc.) that can be used independently in different combinations to support a range of preservation requirements. We will also implement two prototypes that integrate the various technologies developed by the project so as to meet the requirements of the two broad communities involved in the case studies. These prototypes will serve as test beds for the evaluation of the project against the two case studies, and as demonstrators of the project technologies to a wider audience.

4.2 Core research activities

The research in PERICLES is focused around three core research work packages:

- WP3 will develop a conceptual framework and unified model, based on linked data principles, for representing dynamic preservation ecosystems composed of distributed interdependent resources, together with a language and tools describing and managing change in such ecosystems.
- WP4 will investigate and develop a range of analytical techniques and tools for identifying, extracting, analysing and encapsulating information about digital objects and their environments of relevance to their preservation, appraisal and reuse, such as representation information, provenance, contextual information, semantic content descriptions, and metadata more generally.
- WP5 will extend existing lifecycle and preservation models, which focus on technological change, to address the broader evolution of preservation ecosystems, including changes in semantics and user communities, as well in the policies, processes and systems of the preservation infrastructure itself. It will also develop processes and tools that support the management of preservation ecosystems in accordance with these models, and in particular for appraisal processes.

These three work packages are closely interlinked, as illustrated in Figure 1.

- The processual models - and corresponding processes and policies - developed in WP5 will describe and influence the evolution of the more generic models developed in WP3. At the same time, components, such as processes and policies, created by WP5 will themselves be part of the ecosystem and will thus be represented in the model.
- WP5 will produce concrete processes and policies that will be composed from a variety of smaller components, including many of the tools and services for extracting,

analysing and encapsulating information developed by WP4. At the same time, these components will provide relevant metadata to control preservation management and appraisal processes developed by WP5.

- Finally, the information captured and extracted by the tools developed in WP4 will serve to instantiate and populate the linked resources ecosystem model developed by WP3.

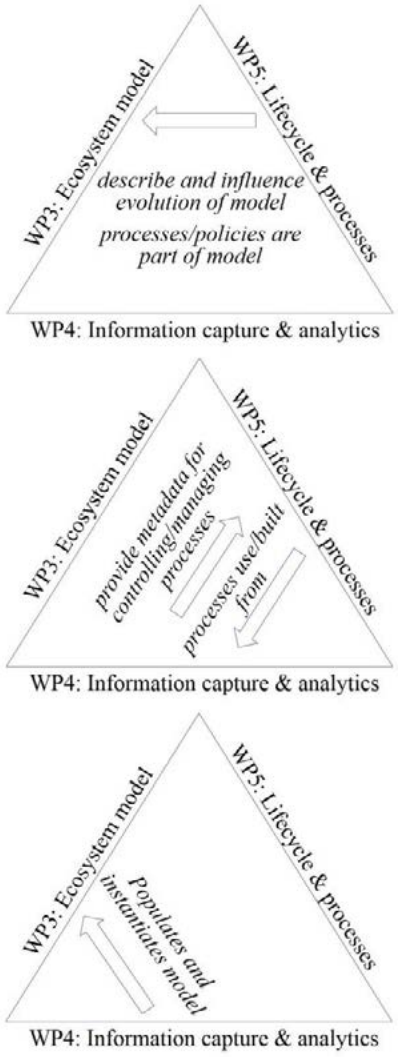


Figure 1. Interactions between research work packages

4.3 Dissemination, communities of practice and exploitation

As well as the two specific communities addressed in the Case Studies, PERICLES will undertake dissemination and engagement activities within a number of communities of practice. Some of these communities will correspond with broad application domains, and will represent an opening out of the domains addressed by the case studies (science and engineering; media and art; archives and other memory institutions); others will correspond to "enabler" stakeholder groups, orthogonal to the domain-based communities and cutting across disciplines (facilities and operations centres; data infrastructure technology; policies and standards; business and sustainability).

The principle behind these communities of practice is that they will function as coordination points for promoting the findings of the project, for seeking input and feedback, and for extending collaborations to new communities. Each community will carry out a range of activities, such as workshops and liaison with external organisations, of relevance to the area it addresses.

A key objective of PERICLES is to set up pathways for the take-up and exploitation of project outputs in production environments, both commercially and in the public sector. While this objective cuts across the communities of practice, because of its importance this will be organised through a dedicated work package (WP10) that will be tasked with identifying exploitation opportunities and developing an exploitation strategy.

5. USE CASES

5.1 Science case study

Preserving space science data is critical for the wider research community. Collecting data in space is extremely expensive - the design of the payload (the experiment device) for operating in orbit is very complex, the cost of launching them to orbit is very high, and operating the payload is very demanding. Moreover, observational data (e.g. sun or weather observations) are simply impossible to replicate.

The science case is based on data from space operations created at B.USOC. B.USOC is one of the European distributed operation centres of the International Space Station (ISS) payloads and is, amongst other functions, Facility Responsible Centre for the SOLAR package. SOLAR is a set of instruments measuring the variations of the energy output of the sun in spectral ranges going from the far ultraviolet to the near infrared. This experiment package has been running on the International Space Station since 2008 but has a longer history. The instruments actually belong to a series extending to first designs in 1976. Thus the current data series span most of the space age and constitute an important source of reference on the impact of solar variations on earth's climate and environment for three solar cycles.

The data from space experiments are collected only during the mission, after the payload is launched to space. However, related metadata (e.g. design documentation) is collected from the start of the project, sometimes many years before the launch – during the mission analysis, feasibility, definition, qualification and production phases. For the SOLAR payload, for example, there are over 900 documents from these preparation phases. These documents include specification and design documents, acceptance data packages, test plans and reports, interface control documents, assessment reports, safety data packages, operation manuals, certifications, thermal analysis etc.

Valuable metadata is collected also during missions – such metadata describes how the experiment devices were operated and can explain for example different anomalies in the results. This metadata includes the different operation interface procedures, flight rules and payload regulations, payload data files, minutes of meetings, console logs, flight notes, checklists, daily operation reports, science planning, command schedules etc.

The space science data itself includes telemetry sent from the payload and telecommands sent to the payload. The telemetry includes housekeeping data – this consists of engineering measurements about the state of the payload (e.g. temperature, voltage and current reading), health and status data which include measurements from sensors outside the payload, and science data.

The raw science data must be calibrated. In SOLAR for example, the detector is put in front of a black body in certain temperature, and since the spectrum of a black body in certain temperature is known, the detector can be calibrated. Another example is that during the mission the detector must be further calibrated as it decays. The SOLSPEC detector is calibrated once in 24 hours during sun's visibility window. The calibration is done using a calibration lamp that its spectrum is known. The South Atlantic Anomaly Disturbances is another example. This anomaly causes an increased flux of energetic particles in the south Atlantic region and exposes orbiting satellites to higher than usual levels of radiation. Apparently, these disturbances have to be taken into account when calibrating the results. The raw science data is processed to include the information from the calibration curves, and may later be further processed to include other calculations of the scientists.

The telecommands are structured data sent to payloads during the operations. They may contain control structures for shutting up or starting various modules, as well as uploads of data and scripts.

In addition to the above, auxiliary data is also collected. Most of auxiliary data comes from public sources. For instance, current B.USOC operations related to the SOLAR payload heavily depend on TLE (two-line elements) to predict the position of the ISS and on the ISS attitude timeline (ATL) to predict the orientation of ISS towards the sun. The two external data sources are combined in order to create a full prediction of the upcoming month allowing clear scientific planning and an optimal operations support plan to be created.

5.2 Media case study

The Media case study, led by Tate, will reflect the activities and responsibilities of distinct areas of the organisation.

- Digital art from the main fine art collection includes software-based art, video and audio content, and file-based material such as vector graphics.
- Born-digital material from artists' estates and from institutional records, for example from the archives of galleries.
- Audio Arts collection. This is a rich archive of digitised audio material generated as an audio magazine that was produced and distributed from 1973 to 2006.
- Tate Media Productions. This content comprises unedited footage, edited programmes, and other digital assets generated in the creation of these programmes.

Each of these categories contains digital assets of very high value, which are moreover associated with a great deal of interrelated contextual information, including (for example) information generated during and (implicitly) documenting the process of creation, as well as content generated in social media sites (e.g. blogs, Facebook, Twitter) once an artefact has been exhibited. The lifecycle of an object thus involves a number of high-level processes that may be represented as formal workflows, which will be used to define the use case.

To take a simple example, in the case of digital art, this description will map out the creation and acquisition of these works into the collection, and the subsequent cycles of display, maintenance and preservation or recovery associated with their life within the museum. The PERICLES project will map not only the digital assets that constitute the components of the artworks themselves, but also the rich information that surrounds them and that describes the context in which they exist as their lifecycle progresses.

The use cases will also describe the existing systems and storage infrastructure used to manage and curate the material, and their part in the broader digital and preservation ecosystem within the institution. Importantly, the use cases will also identify the "pressure points" in the existing workflows, and highlight areas where currently there are in place no practices that implement appropriate preservation strategies for these digital objects.

6. CONCLUSIONS

This paper describes the PERICLES project, its objectives and approach, and details its case studies. The conference poster will describe some early findings of the requirements gathering in both the science and media use cases. These will be used to provide some more concrete examples of the research problems being addressed in the project. PERICLES aims to develop several prototypes to validate the concepts and models being defined, and some initial ideas on these will be discussed.

7. ACKNOWLEDGMENTS

This work was supported by the European Commission Seventh Framework Programme under Grant Agreement Number FP7-601138 PERICLES.

8. ADDRESSES OF AUTHORS

1. Centre for e-Research, King's College London, 26-29 Drury Lane, London WC2B 5RL, UK..
2. University of Borås, Swedish School of Library and Information Science, Borås, Sweden.
3. Information Technologies Institute, Centre for Research and Technology Hellas, Thermi-Thessaloniki, Greece.
4. Dotsoft, Kountourioti, 54625 Thessaloniki, Greece.
5. Georg-August-Universität Göttingen, Wilhelmsplatz, 37073 Göttingen, Germany.
6. University of Liverpool, Brownlow Hill, Foundation Building, Liverpool L69 7ZX, UK.
7. Space Applications Services NV, Leuvensesteenweg, 1932 Zaventem, Belgium.
8. XEROX Research Centre Europe, Avenue du President Wilson, Immeuble Le Jade, 93200, France.
9. Edinburgh Parallel Computing Centre, Old College, South Bridge, Edinburgh EH8 9YL, UK.
10. The Board of Trustees of Tate Gallery, Millbank, SWIP 4RG London, UK.
11. B.USOC (Belgian User Support and Operations Centre), Avenue Circulaire 3, 1180 Brussels, Belgium.

On Preparedness of Memory Organizations for Ingesting Data

Juha Lehtonen, Heikki Helin, Kimmo Koivunen, Kuisma Lehtonen
CSC – IT Center for Science
P.O. Box 405 (Keilaranta 14)
FI-02101 Espoo, Finland
+358 9 457 2001
{juha.lehtonen, heikki.helin, kimmo.koivunen, kuisma.lehtonen}@csc.fi

ABSTRACT
The National Digital Library of Finland is an entity, which aims to create a nationally unified structure for contents and services ensuring the effective and high-quality management, dissemination, and especially digital preservation of cultural digital information resources. The National Digital Library's basis is formed by libraries, archives and museums (partner organizations). Because of the diversity of the partner organizations, the digital content to be preserved makes up a very heterogeneous landscape. To find out preparedness of partner organization to join common digital preservation system, we established ten different pilots of preparing and ingesting submission information packages with common specifications. These specifications define technical requirements for the digital objects, their metadata, and package structure to be submitted for digital preservation. In this paper, we show the piloting process, the experiences and the results, believing that these findings might be useful for various organizations involved with digital preservation.

Categories and Subject Descriptors
H 3.7 [**Information Storage and Retrieval**]: Digital Libraries – *Standards.*

General Terms
Experimentation, Standardization.

Keywords
Digital preservation, designated community, submission information package, metadata.

1. INTRODUCTION
The National Digital Library (NDL) of Finland [1, 2] is an entity within the remit of the Ministry of Education and Culture within the Finnish Government, which basis is formed by national libraries, archives and museums (partner organizations). Almost all memory organizations under the Ministry of Education and Culture of Finland are under an obligation to preserve the cultural material in their possession, of which a lot of is in a digital form. Most of this material consists of digitized documents, maps, photographs, newspapers and sound recordings. In the future, this material will be mostly born-digital, which increases the volume of the data. The aim of the NDL is to ensure the effective and high-quality management, dissemination, and a common digital preservation (DP) of cultural and scientific digital information resources. The objectives of the NDL are ensuring the preservation of digital cultural content, ensuring access to and compatibility of content, designing a cost-effective digital preservation system, promoting cooperation between the partner organizations, and building better services with open cooperation and expansion to include a large range of content. Common infrastructure and services will draw the practices of memory organizations closer, reduce the costs and fragmented nature of the systems, and intensify cooperation.

We are currently designing and implementing the NDL's DP system, which will be based on the OAIS reference model [3]. The implementation is divided into two main phases: the preparation and implementation. The preparation phase will ensure that the original bits of the data can be maintained intact and run on modern hardware. A quick launch in the end of 2013 of the bit preservation will ensure that the digital materials in the possession of the partner organizations can be reliably preserved until the DP system becomes fully operational in 2016. In the second phase, the DP system will ensure that the material remains understandable, its information content can be interpreted, and the material can also be used with the software of coming decades. The system will be built to accommodate the increased volume and diversification of content and organizations, as well as the possible development into a DP system for the research data.

NDL has defined specifications for preparing and creating unified Submission Information Packages (SIPs), with, for example, a redefined METS [4] schema and a closed set of acceptable file formats.[1] As we are building common DP system for various memory organizations, the unified structure enables efficient administration of the information on the long term and also enables semantically commensurable information content. In the NDL METS, some originally optional elements and attributes are stated as mandatory or as recommended, or the use has been restricted. From the acceptable file formats, some are recommended formats, which are straightforwardly accepted for preservation, where others are acceptable for transfer, which are migrated to a recommended format before preservation. The file format selections are mainly based on [5].

The preparedness of the partner organizations and the functionality of the specifications needed to be tested in practice, and therefore, the preparation and creation of SIPs were piloted with the partner organizations. This gave a lot of information about the partner organizations needs and the requirements needed for creating digital data according to the specifications. We believe that these findings might be useful for the various organizations involved with digital preservation. In this paper, we present the experiences and overall results of these pilots. In Section 2, we explain the structure of the pilots and collect the pilot experiences of the partner organization. In Section 3, we give the results found in the analysis of the SIPs. Finally, we conclude these pilots in Section 4.

[1] Specifications are available at http://www.kdk.fi/en, but only in Finnish.

2. PILOTS

To understand the preparedness of the partner organizations for SIP preparation and ingestion, ten pilots were established. Eight of the pilots included documenting the findings and practical work of creating SIPs for our DP system, and two of the pilots were fully reporting exercises. The selected partners (three libraries, five archives and two museums) have been involved with designing the NDL specifications, so they already had some background information about the DP system. All the pilots varied depending on the organization and the selected material, but the basic structure of each pilot was the following: The pilot started with a meeting, where the contact persons, timetable, pilot material types, duties and restrictions were agreed. In the first task, the partner organization identified the mandatory (and depending on time resources, also recommended) metadata fields from their back-end systems, with using the NDL's specifications. The partner organizations were supposed to list all the flaws they found from their system or from the specification, and suggest necessary enhancements to be taken into account in the specifications. In the second task, the organizations collected the test material from their systems, create one or several SIPs according to the specifications and submit them to the ingest pilot implementation. The organizations also wrote a report about their SIP creation experiences, such as listing the easy and difficult tasks, the needed changes in organization's processes or systems, improvement suggestions to the NDL's specifications and so on. The third part of the pilot was a task of the DP system designers, where the ingestion process was tested and documented. This included documenting all the found errors in the submitted packages, but also all the flaws found in the ingestion process. The last task of the pilot was to exchange experiences between the partner organizations and DP system designers.

2.1 Experiences of the partners

The partner organizations found the pilots interesting and useful. They experienced that their knowledge regarding various essential standards increased significantly. Further, the pilot gave them a lot of practical and concrete experience about the information packaging for DP. Three partner organizations found the NDL's specifications and packaging guidelines somewhat easy, inspiring them to create an automated process and choose a heterogeneous set of test material. Three other partner organizations found specifications and the required processes more demanding, and they decided to make the packages by hand. Two partners could provide the packages directly from their current systems, but in these cases, various differences were found against NDL's specifications. Some of the organizations found flaws in their current processes, such as digitizing without creating checksums, inconsistency with the documentation of the digitizing chain, or even missing provenance metadata.

The major feedback from the partner organizations was that providing several different types of examples would have been helpful. Some parts in our specifications were still under construction, such as how to present the rights metadata and the provenance information in various cases, or should different identifier definitions be nationally unified somehow. Also, more instructions were needed for the technical metadata and in several details containing controlled vocabularies. These partner organizations' experiences gave a lot of feedback to be analyzed for the work of the DP system, and as a result of these pilots, the NDL's specifications has been updated.

3. VALIDATION OF PACKAGES

According to the NDL's specification, the partner organization submits one or more SIPs in a ZIP archive to the DP system. The ZIP format is used only for the transfer step, making it possible to transfer one or more SIPs at once. In the first phase of the ingestion, the ZIP archive is unzipped and the structure of it is inspected (see Figure 1). Each first-level directory in the ZIP file is a SIP, which requires a valid METS document and a digital signature at the root of each first-level directories. If needed, there may be subfolders for the digital objects, for example for different manifestos of a given digital object. With this structure, each SIP can be validated separately. In the second phase, the digital signature included in the SIP is validated. The purpose of the digital signature is to validate the origin and the integrity of the data. In our final DP system, the ingestion will be terminated, if the SIP structure is incorrect or if it includes an incorrect signature. In the pilots, the inspection was continued to get all possible information (e.g., errors) from the ingestion. In the next step of the ingestion, the METS document is parsed against our METS schema, including all other schemas used inside the METS document (e.g., PREMIS [6]). In the pilots, the mandatory and recommended metadata fields were also inspected by hand, either entirely or by inspecting a few random samples from the METS document. After this, the checksums of the objects are validated, by comparing the calculated checksums of the digital objects and the checksums given in the METS document. In this phase, it is also inspected that all reported objects exist in the SIP and that there are no loose objects, that is, files without a reference from the METS document. The next step is to validate the file formats to ensure, that each file is correctly formed. After this, an inspection report is created and submitted to the partner organization. In various phases in the pilot, the validation was done with a custom Python or Java implementation including several 3rd party open source software, such as OpenSSL [7] for the signature validation or JHOVE2 [8] for the file validation. In the pilot, the report was done by hand, but automated reporting methods will be implemented to the production system. In our DP system, the last ingestion step is to create the Archival Information Package (AIP) from the SIP, but in these pilots, all the received data was removed from the server.

The validation against the METS schema is not fully enough for the METS documents, and various solutions are currently being built for more complex issues, which can be solved with Schematron [9]. Also, if using JHOVE for XML validation by default, it downloads all the required schemas from the internet for the validation, and therefore the process takes a lot of time. To solve this issue, XML catalogs [10] are required, so that the schemas are loaded locally. Also, the first phase of the packaging is a problem for one partner, where only huge movie files are managed, since creating a ZIP archive takes too much time, and it

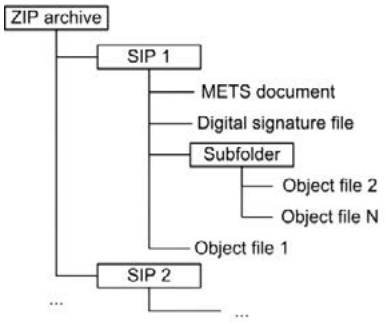


Figure 1. Structure of a ZIP including SIPs.

does not compress the already compressed data. However, this archiving step is used, so that the DP system knows, when the packages have been fully received. The only reasonable solution is to create an exceptional workaround with this partner.

3.1 Overall packaging results

In the analysis of the pilot, a grade between 0-2 was given for each SIP in each validation step as follows:

0. The part is missing or does not follow the specification,
1. The part includes severe errors or a large number of minor mistakes,
2. The part is flawless or includes only a few very minor mistakes.

Since the partner organizations submitted different number of SIPs, the average grade of each step was calculated for each organization separately. The average result of these organization grades for each ingestion step is shown in Figure 2.

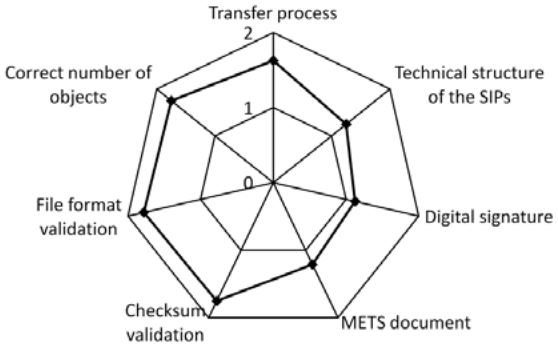


Figure 2. Average grades of the SIPs in validation steps.

From the Figure 2, it can be seen that usually in the SIPs the checksums were calculated correctly, the used file formats were correct and the SIP included the correct number of object files compared to the METS document references. However, the METS document creation had various types of small mistakes (see Section 3.2). The lack of examples generated uncertainty in the details. The creation of the METS document took a lot of time for some partner organizations. It was also found out, that in our specification the technical structure was a little bit confusing, and some of the organizations made various kinds of mistakes in this step, such as got confused, whether the ZIP archive is a SIP or a first-level directory inside a ZIP archive is a SIP.

3.2 Metadata results

In the NDL, a modified version of the METS schema is used, where more specific details have been added in the specification. In the NDL METS profile, the header, descriptive, technical, rights, provenance and struct map metadata are all mandatory, whereas the structural link and behavior metadata sections are forbidden. All the administrative metadata must be placed in a single administrative metadata section, so the use of several administrative metadata sections is denied. However, all the originally mandatory elements and attributes are still mandatory, and all elements and attributes are used in a way that it conforms to the original specification. The original idea was that when using the NDL METS schema, the resulted METS XML file is compatible to the official METS schema. However, as a result of the pilots, few additions were needed to the official METS specification. The request for these additions has been submitted to the METS board [11].

Figure 3 depicts how many organizations had different types of flaws related to the creation of the metadata. The most common mistake related to the creation of the METS document was one or a few missing or misused attributes (a). Five out of eight organizations had some flaws of this type. This is quite expected, since our METS profile includes a lot of mandatory or carefully defined attributes, and one or two of those might be forgotten or misunderstood in the first tryout. The external XML errors (b) were usually small, such as a single misused element or attribute. Namespace issues are quite difficult, and three organizations had problems in that part (cf. (c)). The specification of the provenance and rights metadata was partly under construction, and therefore some of the organizations did not pay attention on those parts (cf. (d) and (h)). Three of the organizations did not submit the digital signature (f), making it impossible to verify that the content of the METS document was correct. For clarity of Figure 3, let us mention that (a) or (n) are error types where one or a few incidental mandatory attributes or elements are missing. Even though for example error type (d) leads to missing attributes and elements, it does not affect to the count of error type (a) and (n), since it already is included in the error type (d).

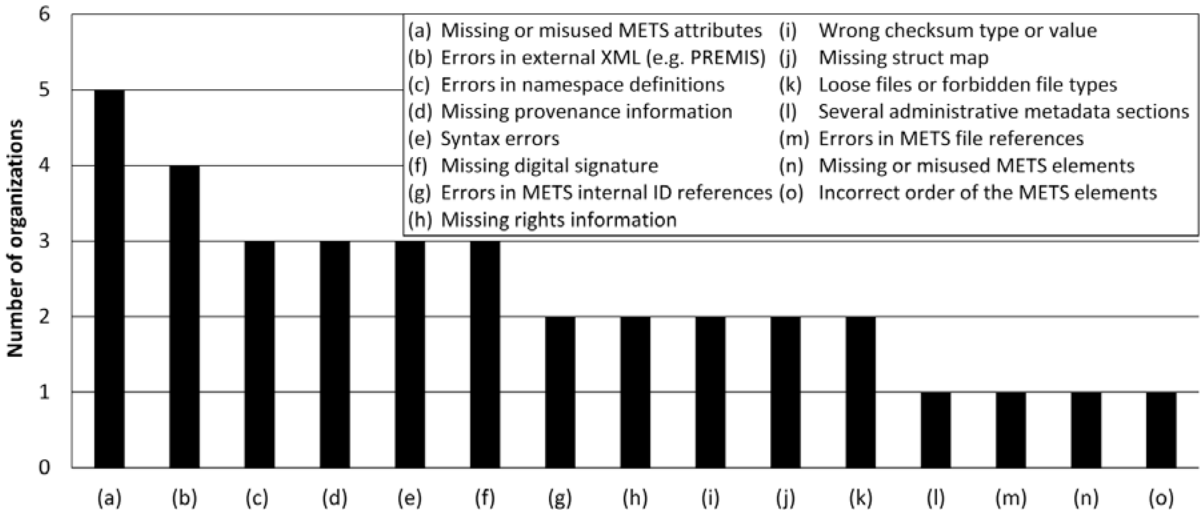


Figure 3. Flaws related to the received METS documents.

3.3 File format results

The file format validation is a process where it is examined whether the file is correctly formed or not. Our approach is to use existing 3rd party software for the validation, as far as possible. The validation was usually done for all the delivered files, but in few pilots, we needed to make the validation by random samples. The results are shown in Figure 4. From these file formats, PDF and MP3 are transferrable formats, where as all others are recommended formats. Most of the files were correctly formed. Some of the PDF/A files included a line feed character, which was a link directing to nowhere, and therefore the ingestion process discarded them. This raised a question, whether the errors of this type are acceptable or not, and how to deal with the issues of this type. We do not have a perfect solution for this, but the current plan is to decide and store a decision value for each error message, which defines a proper follow-up action. The forbidden file formats are file formats not allowed in the DP system, and therefore the ingestion process works correctly when defining those as faulty. The DP system must be built in a modular way, so that if better validation software is found or new file formats are accepted, the validation parts of these formats can be changed or added in a most convenient way.

As depicted in Figure 5, we received mostly JPEG-based files and XML-based files in the pilot. However, what is missing in the Figure 5 is how common certain file types are. Those partners who made the SIPs by hand provided only one or two files in their test packages, where as those partners who created or used an automated process, could provide more files. When JPEG2000 or XHTML files seemed to be quite popular based on Figure 5, but only one organization provided the files in those formats.

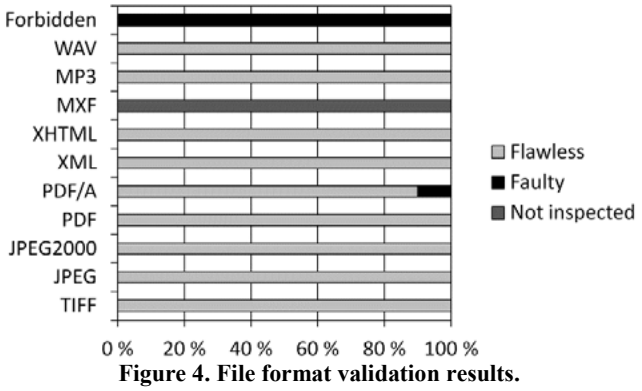


Figure 4. File format validation results.

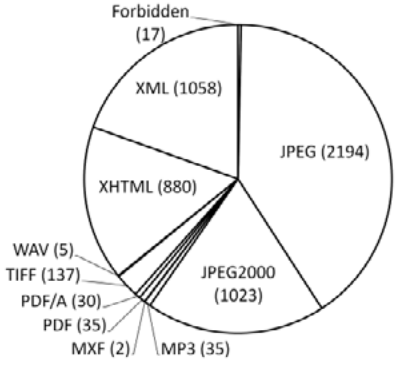


Figure 5. Received number of different file formats.

4. CONCLUSIONS

It was shown that the overall preparedness for preparing digital information for digital preservation is very different in different organizations. The object files are in good condition, but the flaws come up when packaging information and attaching necessary metadata to it. This practically means that the partner organizations are well prepared for a short-term usage of their digital data, but preparedness for a long-term DP needs development. However, it was shown that these problems are not overwhelmingly complicated, and with a carefully designed common technical support system, the partners are able to produce SIPs. One of the major focus point needed is the continuous updating and improvement of the NDL's specifications and operational methods, with paying attention to the required overall workload in package creation.

5. ACKNOWLEDGMENTS

The authors would like to thank all members of the NDL digital preservation support group and technical division for their valuable comments and input during the preparation of the NDL digital preservation system. Further, we thank the Ministry of Education and Culture for funding this project.

6. REFERENCES

[1] *National Digital Library*. URL=http://www.kdk.fi/en

[2] Helin, H., Koivunen, K., Lehtonen J. and Lehtonen K. 2012. Towards Preserving Cultural Heritage of Finland. In *Proceedings of the Cultural Heritage on line – Trusted Digital Repositories & Trusted Professionals* (Florence, Italy, December 10–14, 2012). NBN=http://nbn.depositolegale.it/urn:nbn:it:frd-9299

[3] *ISO 14721:2012: Open Archival Information System – Reference Model*. International Organization for Standardization. 2012.

[4] *METS Metadata Encoding and Transmission Standard*. URL=http://www.loc.gov/standards/mets/

[5] *Local Digital Format Registry (LDFR): File Format Guidelines for Preservation and Long-Term Access, Version 1.0*. Library and Archives Canada. 2010.

[6] *PREMIS Preservation Metadata Maintenance Activity*. URL=http://www.loc.gov/standards/premis/

[7] *OpenSSL – The Open Source Toolkit for SSL/TLS*. URL=http://www.openssl.org/

[8] *JHOVE2*. URL=http://www.jhove2.org/

[9] *ISO/IEC 19757-3:2006: Information technology – Document Schema Definition Language (DSDL) – Part 3: Rule-based validation – Schematron*. 2006.

[10] *XML Catalogs, OASIS Standard V 1.1*. Organization for the Advancement of Structured Information Standards. 2005.

[11] *METS Wiki - Change Requests.* URL=https://github.com/mets/wiki/blob/master/ChangeRequests/NDL-Finland-METS-changes.pdf

Web Archiving as a Service for the Sciences

Anna Kugler
Munich Digitization Center/Digital Library
Bavarian State Library
anna.kugler@bsb-muenchen.de

Tobias Beinert
Munich Digitization Center/Digital Library
Bavarian State Library
tobias.beinert@bsb-muenchen.de

Astrid Schoger
Munich Digitization Center/Digital Library
Bavarian State Library
astrid.schoger@bsb-muenchen.de

ABSTRACT

The collection and archiving of scientifically relevant websites is so far a vastly neglected sphere of activity in German libraries. To counteract this looming loss and providing researchers with permanent access to websites, the Bavarian State Library (BSB) has built up a web archiving workflow already more than two years ago. The main goal of a project newly approved by the German Research Foundation (DFG) is the development and implementation of a cooperative service model. This service will support other cultural heritage institutions in their web archiving activities and facilitate the build up of a distributed German scientific web archive. With this project the Bavarian State Library wants to improve both quantity and quality of scientific web archives and promote their use in the scholarly context.

Keywords

Digital Preservation, Web Archiving, Web Harvesting, Bavarian State Library

1. INTRODUCTION

Digital communication will co-determine the future of the humanities. This statement was beyond all questions at the conference 'Reviewing – Commenting – Blogging: How will Humanities Scholars Communicate in the Digital Future?', which was organized by the Bavarian State Library (http://www.bsb-muenchen.de) at the beginning of this year. The conference was held on the occasion of the 2nd anniversary of the review platform for European History named recensio.net where the participants discussed how far the web as a publication platform can meet the qualitative requirements of scholarship [1]. The presentations and round table discussions clearly showed that new instruments, services and infrastructures need to be developed to publish, evaluate and finally preserve these new types of scholarly resources. One of the new digital resources which have often been "almost completely overlooked", as Nils Brügger put it, "even if they may be considered one of the most significant contemporary contributions to the cultural heritage of mankind" [2] are websites.

Web archives which preserve captures of websites and make them permanently available are still an unknown or unusual type of research instrument for many researchers. Compared to the live web a few distinctive features of web archives however exist, which constitute their necessity:

1. Web archives include content which has already disappeared from the live web. The estimates about the average lifespan of websites differ a lot, but what they show nevertheless is that "although the web can be considered a storage medium of our civilisation, it does not preserve itself for the future – the old web cannot always be found on the web." [3]

2. Already today one concrete use is the possibility to cite websites. Scientists more often refer to or cite online resources, but disappearing content or changing URLs often make consistent access to the cited sources quite difficult or even impossible.

3. Moreover the history of the web illustrates an important part of our culture. Periodic captures of websites not only show the evolution and changing of web technology and web design but also the changing of political and scholarly discourses.

4. In a more technical context it could be possible in the future that certain documents cannot be separated from the tools or platforms which produced them. Based on this fact archiving of websites has a totally different use than archiving of printed books and at the same time makes it much more challenging [4].

5. Last but not least web archives offer a subject-oriented data collection which can be analysed by new types of data mining methods. In the context of the emerging e-humanities scientists can be offered advanced access possibilities.

Nevertheless experience reports of web archives already operating for a long time, such as e.g. the UK Web Archive show that there is still "little evidence of scholarly use" [5]. Next to the fact that many scholars don't know about the existence of web archives, already active users would appreciate an increase of scientific content and see a need for improved data mining tools [6]. Thus the aim of the whole web archiving community, which mainly consists of national libraries and larger regional libraries, has to be to improve quantity and quality of scientific web archives and to promote their use in the scholarly context [7].

2. A DISTRIBUTED GERMAN RESEARCH LIBRARY – RESULTING IN A DISTRIBUTED GERMAN RESEARCH WEB ARCHIVE?

In Germany the state of web archiving activities looks a little bit like a rag rug of small initiatives collecting websites. In some cases the collection strategy derives from a regional legal deposit obligation (e.g. BOA, Baden-Württembergisches Online-Archiv, http://www.boa-bw.de/ or edoweb, digital archive for Rheinland-Pfalz, http://www.lbz-rlp.de/cms/landeskunde/edoweb/), while other regional libraries do not archive websites at all. The German National Library, which is the legal deposit library for Germany, started harvesting the most important websites significant for German society, history, politics, economics and culture last year and intends to make them accessible in their reading rooms [8]. In summary web archiving in German libraries does not fulfil the needs of the scientific community so far. This is partially owed to the federally structured German library landscape.

In Germany due to the political and historical situation one central research library has never been established. In 1949 a special cooperation system was launched which is internationally unique and takes responsibility for the supra-regional literature supply of scholarship and research. It is organised by the German Research Foundation. 23 state and university libraries and some specialised libraries participate in this cooperation system, each of which is responsible for one or more scientific subjects (Sondersammelgebiete, SSGs). All libraries participating in this system pledge themselves to make their collections available nationwide. The access point to the collections is offered by so-called virtual subject libraries which bundle the diverse search possibilities under one user interface („one-stop-shop“) [9].

The collections of these subject libraries include freely accessible internet resources with scientific relevance which are to a great extent websites. Each website is intellectually selected by experts of the respective specialist department and a lot of effort is put into the cataloguing process. Several libraries decided to cooperate concerning the cataloguing system and they built up Academic Link Share (ALS) (http://www.academic-linkshare.de), a database system already containing more than 100,000 entries. Not only the content itself but also this time-consuming work is lost as soon as the respective website disappears, as until now archiving is not mandatory for these subject libraries.

A thorough evaluation process of the system of special subject areas commissioned by the German Research Foundation in 2010 with the aim to find out how the needs of the sciences could be best fulfilled in the future resulted in several recommendations: a future acquisition focus on electronic resources, a stronger focus on customisation to the scientific needs concerning collection building, and in general an improved service-orientation for scholarship. In order to guarantee sustainability, long-term preservation and permanent access to digital data new service models should be developed [10].

In this context BSB applied for a project at the DFG which focuses on developing a cooperative service model for web archiving. It will be built on the infrastructure and experience BSB has already gained in creating a scientific web archive over the last two years and thus support other cultural heritage institutions in their web archiving activities. A closer cooperation with scientific communities is an important aim. The outcome of the project could become the nucleus for a distributed German Research Web Archive. The project started at the beginning of 2013.

3. WEB ARCHIVING AT THE BAVARIAN STATE LIBRARY

In 2010 BSB's Munich Digitization Center/Digital Library (MDZ) (http://www.digitale-sammlungen.de) began to collect and archive websites in a pilot phase, since the beginning of 2012 its web archive increases productively (http://www.babs-muenchen.de). The focus is on websites already selected and catalogued by the virtual subject libraries (Virtuelle Fachbibliotheken, ViFas). BSB e.g. is responsible for five virtual subject libraries dealing with the fields of:

- History (www.propylaeum.de, www.historicum.net)
- Musicology (www.vifamusik.de)
- Eastern Europe (www.vifaost.de)
- Romanic culture area (www.vifarom.de)
- Library, book and information studies (www.b2i.de)

These virtual subject libraries account for about 10,000 entries of websites in the ALS database. As Bavaria's existing legal deposit regulations only allow to harvest and archive websites of Bavarian authorities but not scholarly websites in an international context an explicit permission is necessary to harvest, archive and make accessible those websites. Thus a very detailed permission request (email) is sent to each 'website owner' in which the rights to harvest and archive the website in regular sequences and to make it available on the web are requested if no rights of third parties interfere. Moreover in the permission request email it is pointed out that German copyright law applies. The positive return rate is about 20% to 30%.

For the harvesting process BSB uses the Web Curator Tool (WCT), an open source software developed jointly by the British Library and the National Library of New Zealand. It was chosen because it allows an integrated process for selective web harvesting including the administration of the permission request, harvesting with job scheduling and a partly automated quality control. The website crawler is Heritrix which has been developed by the Internet Archive. To provide access to the crawled websites a local adaptation of the Wayback Machine has been implemented. The harvested WARC (Web ARChive) files are archived in Rosetta, a commercial software for digital long-term preservation by Ex Libris. Ensuring bitstream preservation is done by the Leibniz Supercomputing Centre (LRZ), whom the BSB has been working together closely for many years now.

All harvested and archived websites are freely accessible. The harvested websites are made available firstly via the index of the virtual subject libraries and secondly via our local library catalogue. The first access point to the archived websites is based on the indexed metadata of the virtual subject libraries. For each internet resource there is a detailed ALS entry containing title, URL, responsible institution, contact, key words and even a short abstract. Next to the link to the live website there has been added a so-called „Archive-URL“, which is a stable link leading to the archiving system Rosetta, where the websites are preserved. If the user clicks on this “Archive-URL” link a local adaptation of the Wayback Machine opens and shows all the archived captures for a single website. For citation purposes persistent identifiers (Uniform Resource Names, URNs) will be assigned on website level or even on capture level. In BSB’s catalogue system each website is described by a minimal catalogue entry. About 500 archived websites (with several captures) can be found in BSB’s catalogue system so far (June 2013).

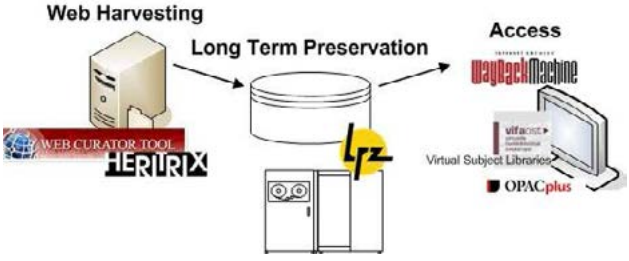


Figure 1: BSB’s Web Archiving Workflow

4. BUILDING UP A SERVICE FOR OTHER CULTURAL HERITAGE INSTITUTIONS

The newly approved DFG project focuses on creating a web archiving service for other (German) cultural heritage institutions like e.g. the Web Archiving Service (WAS) of the Californian Digital Library (http://webarchives.cdlib.org) or Archive-It from the Internet Archive (http://www.archive-it.org). Building up this service is accompanied and based on several improvements and extensions of the existing infrastructure.

The first work package of the project aims at developing a collection and archiving profile for websites which will constitute a very important guideline for the institutions using the new web archiving service. The collection profile will rely a lot on the collection criteria for scientific internet resources already defined by DFG, but additionally include criteria which derive from BSB’s specific collection policy as archiving library of Bavaria. Very challenging is the definition of an archiving profile which has to address content criteria as well as technical and temporal coherence aspects. Due to technical challenges the available harvesters often are not able to crawl an identical mirror of a website, especially dynamic functionalities like e.g. database queries that won’t work any longer in a web archive. Maybe these URLs need to be excluded from the archiving process although the content is highly valued due to the collection profile. Moreover the crawl duration and frequency is of course critical referring to completeness and temporal coherence of a harvested website [11]. Thus an archived website can always only be a “reconstruction” [2] of the live website, which needs to be reflected in the archiving profile. A detailed collection and archiving profile will help to design the defined crawl scope more exactly and to improve the quality of the web archive.

Improving the quality assurance process and defining quality criteria is another task in the new project, which still poses a challenge for many institutions in the web archiving context [12]. At BSB the quality of a new website capture is still evaluated manually by visually comparing the live website with the archived version and by analysing the log files of the crawler. But the increasing amount of website captures makes the development of more automated workflows inevitable. Thus defining quality criteria and testing new tools and methods already used by other institutions and recommended by the International Internet Preservation Consortium (IIPC, http://www.netpreserve.de) will help to develop a (partly) automated workflow. Moreover the new version 1.6 of the WCT offers a partially automated workflow for quality assurance on the basis of technically reviewable quality indicators, which will be implemented as soon as possible.

Also in terms of access new ways have to be explored, as a mere searching for URLs or titles of websites in many cases no longer fulfils researchers’ needs. [13]. Therefore the already existing data mining technologies, that might be appropriate for web archives have to be analysed and tested. A special focus is to be put on full text indexing and search, with SOLR being considered as a possible solution [14]. Another way to improve the accessibility of the content of the web archive could be to work on Memento compatibility, so that historical versions of websites from different web archives can be integrated into the live web via a simple browser extension [15].

Another work package deals with available long-term preservation measures for web archives. Different tools and methods for the use of digital long-term preservation of websites will be tested and probably implemented like e.g. JHOVE2 for the format characterisation of the files inside the WARC files. On the basis of this format characterisation several format migration tests will be pursued. Moreover deduplication tests are part of this work package in order to gain more knowledge about the possibilities of storage reduction processes with the Tivoli Storage Manager (TSM) which is used at the LRZ or already at harvesting time with the latest Heritrix version. The preservation system Rosetta offers certain risk assessment functionalities and format migration solutions which will be tested for websites during the project.

Based on these experiences and improvements the conceptual phase will start for building up a cooperative service model for web archiving for other (German) cultural heritage institutions. To achieve this BSB will work together with external partners, find out about their requirements for web archiving and design a basic service concept with different service levels. First of all the service levels cover the selection and harvesting process which could be done centralised at BSB or decentralised with a complete new software installation at the partner institution. Secondly the archiving and preservation responsibilities need to be discussed which will be most probably solved best centralised at BSB. The third service level deals with access, which depends very much on the requirements of the partner institution. The most crucial point

is whether free access is possible or not and which search possibilities are favoured.

After the conceptual work is done, a technical implementation of the service model has to be realized. That includes most probably further installations of the WCT, setting up the required interfaces and possibly own technical extensions. In a final step after an intensive testing phase the overall costs for a cooperative web archiving routine have to be precisely calculated and put into a business model.

5. CONCLUSION

The ambitious work programme described above aims at improving the state of web archiving in Germany. With an extended and cooperative infrastructure the already existing selection and cataloguing capacities can be integrated in a much more sustainable process that includes an archival component. Thus enduring access to scientifically relevant websites for researchers can hopefully become the norm rather than the exception.

6. REFERENCES

[1] http://rkb.hypotheses.org/410and http://www.ahf-muenchen.de/Tagungsberichte/Berichte/pdf/2013/038-13.pdf

[2] Brügger, N. & Finnemann, N. O. 2013. The Web and Digital Humanities: Theoretical and Methodological Concerns. In: Journal of Broadcasting & Electronic Media 57,1 (2013), p. 66-80, here p. 79.

[3] Brügger, N. 2012. Web History and the Web as a Historical Source. In: ZeithistorischeForschungen / Studies in Contemporary History, Online Ausgabe, 9 (2012), H. 2: http://www.zeithistorische-forschungen.de/16126041-Bruegger-2-2012.

[4] van den Heuvel, Ch. 2010. Web Archiving in Research and Historical Global Collaboratories. In: Brügger, Niels (ed.): Web History. New York 2010, p. 279-303.

[5] Hockx-Yu, H. 2013. Scholarly Use of Web Archives: http://files.dnb.de/nestor/veranstaltungen/2013-02-27-scholarly-use-of-web-archives_public.pdf.

[6] Meyer, E., Thomas, A. & Schroeder, R. 2011. Web Archives: The Future(s): http://netpreserve.org/resources/web-archives-futures

[7] Meyer, E. 2010. Researcher Engagement with Web Archives: Challenges and Opportunities: http://repository.jisc.ac.uk/543/1/JISC%2DREWA_ChallengesandOpportunities_August2010.pdf

[8] Cremer, M. 2013. Providing Access to the DNB Web Archive: http://files.dnb.de/nestor/veranstaltungen/2013-02-27-providing-access-to-the-DNB-web-archive.pdf

[9] http://webis.sub.uni-hamburg.de/webis/index.php/Wissenschaftliche_Bibliotheken

[10] http://www.dfg.de/download/pdf/dfg_im_profil/evaluation_statistik/programm_evaluation/studie_evaluierung_sondersammelgebiete_empfehlungen.pdf

[11] Spaniol, M., Mazeika, A., Denev, D. & Weikum, G. 2009. Catch me if you can: Visual Analysis of Coherence Defects in Web Archiving. In: The 9th International Web Archiving Workshop (IWAW 2009). Workshop Proceedings, p. 27-38: http://www.iwaw.net/09/IWAW2009.pdf

[12] Gray, G. & Scott, M. 2013. Choosing a Sustainable Web Archiving Method: A Comparison of Capture Quality. In: D-Lib Magazine 19, Number 5/6 (2013): http://dx.doi.org/10.1045/may2013-gray

[13] Niu, J. 2012. Functionalities of Web Archives. In: D-Lib Magazine 18, Number 3/4 (2012): http://dx.doi.org/10.1045/march2012-niu2

[14] Pennock, M. 2013. Web-Archiving: DPC Technology Watch Report 13-01 March 2013, p. 23: http://dx.doi.org/10.7207/twr13-01

[15] http://www.webarchive.org.uk/ukwa/info/mementos

A Collaboration to Clarify the Costs of Curation – The 4C Project

Neil Grindley
Jisc
London
UK
+44 (0)203 006 6059
n.grindley@jisc.ac.uk

ABSTRACT

This poster will describe actions being taken in the context of the 4C Project (a Collaboration to Clarify the Costs of Curation), an EC-funded two-year coordination action that aims to promote a better understanding of the potential for undertaking digital curation activity. The approach it is taking is to focus firstly on costs but then to link that concept to related ones such as benefits, value, risk and sustainability, therefore taking a holistic economic view of digital curation. This is important as it links up with various strands of previous work, both on costs and activity models; and on benefits, sustainability and the broader economic framework for digital preservation and access. The purpose of the poster is to describe the novel framework for activity that was proposed in response to the EC FP7-ICT-2011-9 call, and to summarily describe some of the outputs and objectives in a graphical and accessible format.

Keywords

digital curation, costs, economics, benefits, sustainability, risk, 4C, European Commission

1. INTRODUCTION

The European Commission's FP7 ICT 9 programme [1] invited proposals that would be "promotion schemes for the uptake of digital preservation research outcomes including outreach to new stakeholders and roadmapping activities." On the face of it, this meant coordination actions that would synthesise existing work and improve its uptake and implementation across a range of different communities. Underpinning this, however (and articulated at the briefing meeting organised by the EC to describe the aims of the call), was a sense that - despite significant investment - not enough tangible progress had been made with devising workable and competitive solutions and services in the digital preservation realm.

It was suggested that what was required was a healthy and diverse market for mature technical solutions that tackled the real long-term digital asset management problems that all types of organisations face on a daily basis. The conclusion seemed to be that if this could be orchestrated, then the consequent supply and demand would provoke the sort of activity - particularly the flow of services and solutions from SME's (small to medium commercial entities) towards institutions - that was so urgently required at macro-economic European level.

The challenge was, therefore, to design a project that usefully synthesized an area of digital preservation research; was of wide interest to a variety of stakeholders in different working domains; was capable of representing and enhancing that existing work; and tackled the topic in such a way that it would shed useful light on the barriers to uptake and the implementation of solutions and services.

The answer alighted upon by a consortium of partners brought together under the banner of the 4C project was to focus on the costs and economics of digital preservation. The particular challenge thereafter was to devise the best way of drawing together a substantial and heterogeneous body of work; to enhance and then present it afresh to new stakeholders; and finally to make it easier for future actors in the domain to either demand or supply digital preservation solutions and services.

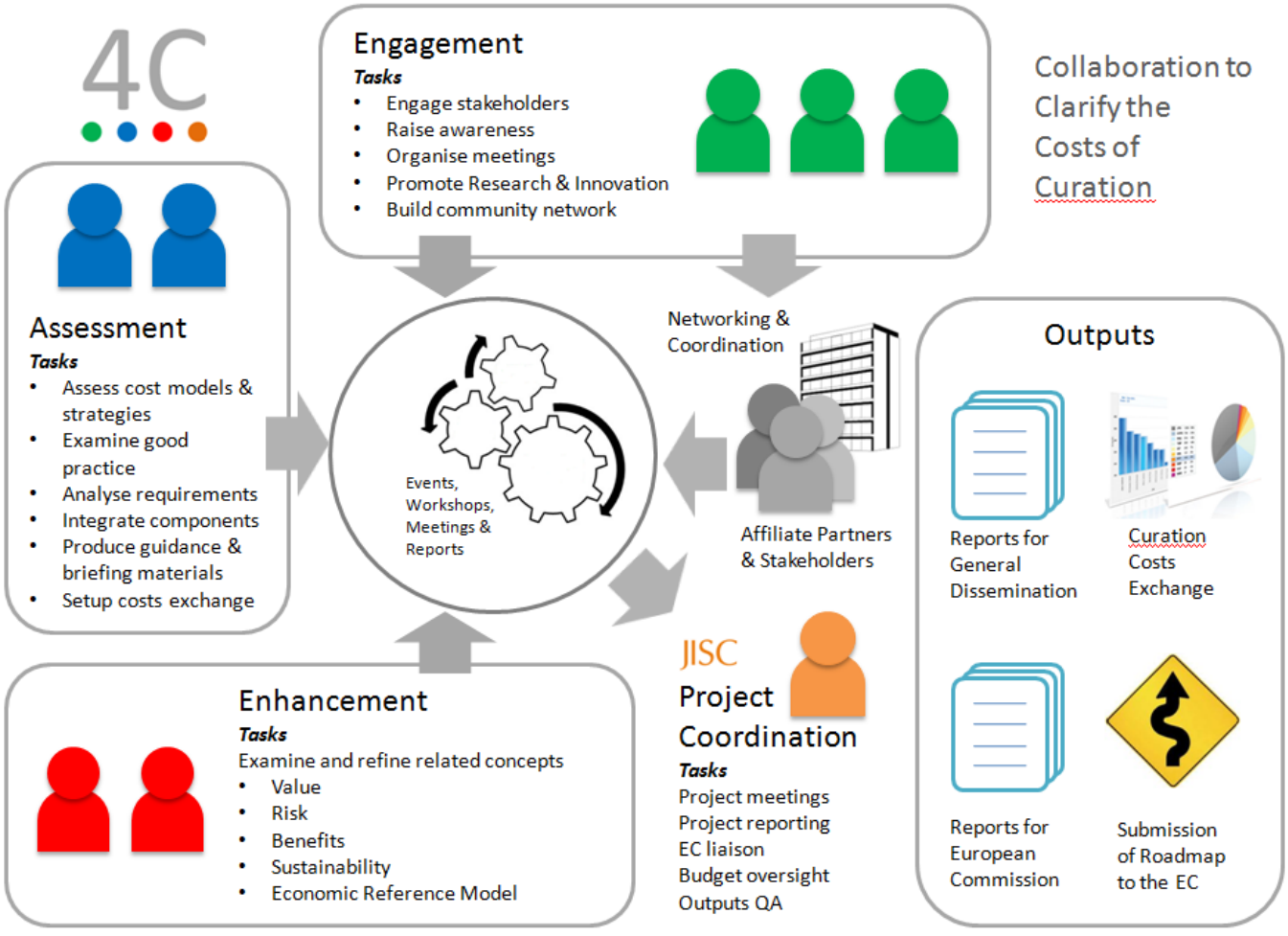
2. EARLY ASSUMPTIONS

There is, as stated above, a substantial amount of work that already exists on the topic of the costs and economics of digital preservation and curation. This can be found in bibliographies [2] and in project listings [3] and includes initiatives that have: formulated cost models (e.g. LIFE project, CMDP project); suggested frameworks (e.g. Keeping Research Data Safe); formed task forces (e.g. Blue Ribbon Task Force on Sustainable Digital Preservation and Access - BRTF); written reports (e.g. APARSEN); and doubtless produced other types of output over nearly twenty years of activity. It was clear at the outset that the 4C project did not need (and should not try) to formulate a cost model to surpass and encapsulate all other cost models. Even if it were possible within the constraints of an EC-funded coordination action, it was apparent that what was required was to build on, join up and communicate this wealth of existing work rather than do new development work and risk duplicating prior effort.

What was also apparent was that the cost of digital *curation* (a term used interchangeably with *preservation* and *archiving* for the purposes of this project) was not a concept that could exist in isolation from a whole raft of other issues. Taking the broader economic view, the 4C Project classes digital curation as an investment, and whilst there are costs associated with an investment, the point is to realise a return or a benefit. Understanding what sort of ratio of cost to benefit organisations will be able to realise (and over what timescales) starts to unpack the whole complexity of digital curation and necessitates examination of other issues such as the level of risk that organisations are willing to accept and other issues such as the value that they attach to those assets. Additional factors and issues that might affect the cost of digital curation could include: trustworthiness, quality, sensitivity, confidentiality, authenticity, capability etc. These factors are referred to as *indirect economic determinants*.

There was also an assumption that the 4C project needed to build a firm foundation for future work and to set an agenda. It was, therefore, agreed that although it did not seek to build another detailed cost model, it could reasonably aspire to: assemble conceptual models; define generic specifications; and to provide a platform for collecting information that ought to be of common interest to a broad range of organisations using a mechanism called the Curation Costs Exchange (CCEx). These form some of the deliverables of the project and will be described in more detail in the poster. Some of them are referenced in figure 1 below.

Figure 1: Conceptual Structure and Mechanism of the 4C Project



3. PROJECT ELEMENTS

The project structure was purposefully conceived as comprising of a minimal set of entities. The complexity of the conceptual issues and the difficulties around terminology were felt to be severe enough without introducing disparate and numerous working groups. The project was therefore formed into five work packages, the first of which was a standard component of project management. An articulation of the other four work packages usefully outlines the purpose and the internal dynamic of the project, also set out in figure 1, which graphically illustrates the project mechanism. (This image will feature as the core of the poster. Further graphical elements will surround the core and will elaborate on the individual components and the approaches taken.)

3.1 Engagement

Engagement is the key activity of the 4C project and informs the entirety of the rest of the work. In line with its coordination action status, it is the principal purpose of the project and will determine whether or not the initiative makes a positive and lasting impact.

The main objective of the Engagement group is to engage with a wide range of stakeholders in memory institutions, universities, SME's, government, data intensive research, industry etc. It will identify, get involved and build partnerships with individuals, groups and institutions that are active or interested in the issue of curation costs and it will attempt to foster a better understanding of the issue amongst the community more broadly. This will be done using well-rehearsed outreach techniques but will also be facilitated by the analysis, expertise and outputs from the other two main project groups. It is this aspect of the project that provides the necessary enhancement on existing work and should enable the discussions to be threaded through with a more

sophisticated understanding of the complex conceptual relationships involved with cost issues.

3.2 Assessment

The main objective of the Assessment group is to establish the most effective current methods for private and public sector organisations to estimate and compare the cost of digital curation, and to identify the most beneficial paths for future development of solutions and services. This will enable stakeholders to more effectively and comprehensively assess the investment of resources that may be required to sustain their digital preservation activities; and allow comparisons of existing and future tools and models with the knowledge that a broad range of criteria: e.g. price, savings, quality, value, risks, benefits, sustainability, etc., are implicit to the comparison.

3.3 Enhancement

One of the key objectives of this group is to ensure that comprehensive and appropriate consideration is given to all indirect factors that might be considered economic determinants of digital curation. Whilst the Assessment group is trying to harmonise and synthesise, the purpose of the Enhancement group is to entertain complexity, consider broader conceptual issues, and to worry about how and when indirect factors should feature in the organizational planning of stakeholder organisations.

Two key factors identified at the outset by the project are:

- Levels of trustworthiness aspired to by organisations and consequent activities around certification
- The level of risk an organisation is prepared to accept

Other economic determinants will emerge as more or less of a priority in the course of the work and as part of the engagement dialogue with the community.

The Enhancement work is also driven by two other key imperatives. The first focuses on sustainability and builds on existing work [4] to develop a draft Economic Sustainability Reference Model (ESRM). This will elucidate the threats to digital assets, the timing of those threats, and help to define terminologies.

The other imperative is to move from the area of costs and into the realm of business models, which addresses the concerns raised by the European Commission at the outset (see above) and also segues into the last of the work packages which is to look forward and define an agenda for research, development and collaboration.

3.4 Roadmap

The purpose of this activity is to arrive at coherent and evidence-based recommendations for future action and strategy in relation to the economic aspects of digital curation. The focus will be on measures that will assist diverse types of organisations to better understand and take control of the cost of managing digital assets over varied timescales, including the provision of cost-effective solutions and services to others. This roadmap report will synthesise and exploit the valuable intelligence that emerges from the other work packages and will also ensure that the content and conclusions are complementary and non-duplicative of work being taken forward by others.

4. PROJECT OBJECTIVES

As should be clear from the preceding information, the project has a number of different objectives. Perhaps the most practical and immediate is to provide organisations that have a need to curate data with more effective and accurate ways of working out how much this activity will cost them. This will help them to do more effective planning and resource allocation.

This first objective is straightforward in cases where the principal business of an organisation (e.g. a national library or archive) is curation. However, this is not the case for most organisations, so the purpose of clarifying the cost of curation should also serve to bring into sharper relief the reasons why curation should or shouldn't be resourced. Therefore, the 4C work will engage not only with the costs of curation but also the benefits that it might realise. Or to put it another way - from an economic perspective - it will examine curation as an investment and be mindful that investments require returns, involve elements of risk, and connect with notions of sustainability and business planning.

The second principal objective is to synthesise, make sense, re-present - and where appropriate enhance - the previous and emerging valuable work that has been done in this area over the years. The costs, economics and sustainability of curating digital assets has been tackled from many perspectives by initiatives across the world but is still not widely understood or effectively embedded into practice. 4C will therefore undertake advocacy and promote relevant work to existing and new stakeholders.

A third important objective is to try and help address the underperforming market place for digital curation solutions and services. One practical way that 4C can assist with this is to establish more effective and accurate ways of predicting the cost of curating materials. If more accurate costs can be relied upon, then more confident designs can be produced for services and solutions (e.g. third party archiving services) that can either run at a profit in the commercial realm, or be assured of breaking even if in the not-for-profit sector.

5. POSTER OBJECTIVES

This poster will address a number of objectives:

- It is a publicity and dissemination opportunity for the 4C project
- It is an invitation to stakeholders to engage with the project and to identify with its aims and objectives
- It sets out a concise description of a project with complex objectives
- It tests out some assumptions and approaches which will require broad community acceptance and endorsement if the project is to be influential and have a positive impact

6. ACKNOWLEDGMENTS

Some of the content of this proposal has been assembled from the input of other 4C project partners. The list of organisations who are participants in the project is as follows:

1. Jisc (UK)
2. Royal Danish National Library (DK)
3. INESC-ID (PT)
4. Danish National Archives (DK)

5. German National Library (DE)
6. University of Glasgow – DCC (UK)
7. University of Essex – UKDA (UK)
8. KEEP Solutions (PT)
9. Digital Preservation Coalition (UK)
10. Secure Business Austria (AT)
11. University of Edinburgh - DCC (UK)
12. Data Archiving & Networked Services (NL)
13. National Library of Estonia (EE)

7. REFERENCES

[1] European Commission participant portal (2011), http://bit.ly/15PNtnA (accessed 27/04/2013)

[2] Lazorchak, B. 2012. A Digital Asset Sustainability and Preservation Cost Bibliography, Library of Congress, http://1.usa.gov/RYMZW2 (accessed 27/04/2013)

[3] Wheatley, P. (and other wiki contributors) (2013). Digital Preservation and Data Curation Costing and Cost Modelling, http://wiki.opf-labs.org/display/CDP/Home (accessed 27/04/2013)

[4] Lavoie, B., Rusbridge, C. (2011). Introduction to Economic Sustainability Reference Model, Blog, http://unsustainableideas.wordpress.com/2011/10/06/intro-reference-3/ (accessed 27/04/2013)

TAP: A Tiered Preservation Model for Digital Resources

Umar Qasim
University of Alberta
Alberta, Canada
umar.qasim@ualberta.ca

Sharon Farnel
University of Alberta
Alberta, Canada
sharon.farnel@ualberta.ca

John Huck
University of Alberta
Alberta, Canada
john.huck@ualberta.ca

ABSTRACT

Rapid changes in the field of technology and exponential increase in the volume of digital content makes long-term preservation of institutional resources a challenging task. Digital preservation requires a commitment for applying preservation actions along with continuous monitoring and management of the preserved resources. The expense of these actions mean that a memory institution needs to make choices about what level of preservation it can afford to provide for a resource when it makes a commitment to preserve it. This paper presents a tiered model to determine preservation levels for digital content, based on an assessment that considers three factors: type of resource, archival responsibility, and projected preservability of the resource (TAP). The paper presents a practical, flexible approach to a complex set of factors and includes examples of how the model can be applied at an academic library.

1. INTRODUCTION

Technological obsolescence is a well known phenomenon and organizations require enormous amounts of resources, both human and financial, to deal with this challenge. This issue becomes even more challenging for memory institutions which are dealing with a wide range of digital resources. Given this situation, common strategies used for preservation, such as emulation, normalization, and migration, may become very expensive to apply across the board.

In this paper we present a tiered assessment model for preserving digital resources at memory institutions. The TAP model assesses digital resources based on three factors: **T**ype of resource, **A**rchival responsibility, and projected **P**reservability level. Institutions can use this model to separate digital resources of enduring value that require rigorous preservation actions from those that require only minimal preservation operations and are intended to be preserved for a short period of time. The model is described in the section 2 and an implementation of the model is discussed in the section 3.

2. THE TAP MODEL

Digital preservation requires a set of processes and activities to ensure long term access to digital resources but do not require the same strategies for every single object. In some cases, resources might only need to be preserved for a short time period, whereas medium and long term preservation may only be needed for some specific resources. The tiered preservation model helps in assessming resources and is based on three factors: type of resource, archival responsibility, and projected preservability, as detailed below.

2.1 Type of Resource

The first evaluation factor, type of resource, considers the nature of the resource from a variety of perspectives, and bears similarities to acquisition or digitization selection policies. In fact, preservation selection criteria rest on the foundation of acquisition and digitization selection policies [14]. This is especially true when an institution is primarily acquiring digital resources [3][2]. However, other factors also merit consideration when selecting for preservation. An institution will wish to safeguard the investment it has already made in a resource [6][4]. Institutions are often stewards of digital resources acquired or created through diverse means, beyond local digitization, and that range must be taken into account [10]. When institutions hold unique material of enduring value, they have a special relationship to that material, as it unlikely to be preserved elsewhere [13][15]. In particular, we suggest a set of five resource types with different scores as shown in table 1.

The first type of resource is Collections of strength. These are resources that the institution has designated as signature collections according to internal defined criteria. These types of resources are promoted at a strategic level and reflect the identity and reputation of the institution. They are the result of a significant investment in time and money, and their content is significant and unique. They may be flagship digitization projects based on special collections holdings or the research focus of the parent institution.

The second type of resource is Locally created, born digital resources. These are resources that are comprised of unique content created in the context of the parent institution's core activities. They represent a significant investment by the institution, and would not necessarily be preserved elsewhere, but lack the profile or focus to be a Collection of strength. An example is a campus institutional repository.

The third type of resource is Other locally digitized or purchased resources. These are resources that the institution has digitized or has had digitized, and therefore owns, but which are not necessarily unique holdings or closely related to core mission. Digitization may have been a result of convenient opportunity. Retrospective scanning of microfilm series or newspapers are examples.

The fourth type of resource is Licensed resources with perpetual access rights. These are resources that the institution has invested funds in to ensure perpetual access, but which it does not own or bear exclusive responsibility for. They

Table 1: Types of resources

Type of Resource	Score
Collection of Strength	5
Local Born Digital Resources	4
Purchased / Digitized Resources	3
Licensed Resources	2
External Resources	1

Table 2: Scoring Levels for Archival Responsibility

Archival Responsibility	Score
Sole Responsibility	2
Shared Responsibility	1
Third-party Responsibility	0

may be key resources that are heavily used by local users.

The fifth type of resource is Externally created, digital resources that are of great value and significance to the institution. These are resources that the institution has assumed stewardship of, though they originated elsewhere. Responsibility to preserve these resources may be the result of strategic decisions made by the institution or its parent organization. An example is an at-risk collection of digital resources created in the local community.

2.2 Archival Responsibility

The number and types of resources that are either born digital or digitized is vast and continues to grow at an increasing rate. For this reason, memory institutions have for some time understood that no single organization can be responsible for preserving them all, nor can, or should, any memory institution preserve its own digital content without engaging in collaborations and partnerships [17][19][8][7]. In our model, we use three types of archival settings as described below.

The first category of archival responsibility is sole, which indicates that the resource is being preserved only by the institution itself. An example may be locally digitized content. The second category of archival responsibility is shared, which indicates that an institution is engaged in a collaborative preservation effort. An example might be Open Journal System content preserved as part of a LOCKSS network. The third category of archival responsibility is third-party responsibility, which indicates that an institution has determined that a third party is more suitable for ensuring the long term accessibility of a digital resource, and so has outsourced preservation responsibilities. An example might be partner resources digitized and available through the Internet Archive. Table 2 shows scores for different categories of archival responsibility.

2.3 Projected Preservability

Projected preservability is a measure to determine the likelihood that a digital resource will be accessible and usable in the long run. Resources at a higher level of projected preservability indicate a higher degree of confidence in providing preservation commitments and are more likely to be accessible in the future. Researchers and practitioners have identified a number of factors that can help to project the preservability of a file format or in other words to determine the level of projected preservability of a resource. TAP model uses five different determinants, i.e. adoption, openness, transparency, stability and interoperability to measure the projected preservability of a resource as discussed below.

2.3.1 Adoption

Adoption is the extent to which a file format has been widely adopted and formally selected for preservation by memory institutions [18]. This information is captured from other memory institutions' published resources when their local registry of file formats is publicly available. Low adoption means no one else is using this file format for preservation, medium adoption is if less than 50% of the recorded institutions are recommending this file format for preservation and high means 50% or more of the recorded institutions are recommending this file format for preservation.

2.3.2 Openness

Openness is the extent to which a file format specification is in the public domain [16][9]. An open file format has a published specification for encoding information, usually maintained by a standards organization, and can be used and implemented by anyone. Open file formats are expected to have less chance of being locked in by a specific technology and/or vendor than proprietary formats. Since the specifications are known and open, other institutions are likely to implement the same solution adhering to the same standard. Hence, openness offers better protection of the digital files against obsolescence of their applications. Proprietary file formats are considered at a low level of openness, whereas Non-proprietary file formats are considered at a medium level and non-proprietary and standardized file formats are considered at a high level of openness.

2.3.3 Transparency

Transparency is the extent to which the contents of a file are open to the direct analysis using basic tools such as, human readable text editors [18]. Additionally, audio/video file formats concealed with compression and wrappers are less transparent and prone to higher preservation complexities. Both of these characteristics, human readability and compression, indicate how complicated a file format can be to decipher. If a lot of effort has to be put into deciphering a format, and with the chance it will not completely be understood, the format can represent a danger to digital preservation and long-term accessibility. Textual file formats which use simple and direct representation will be easier to migrate to new formats and are preservation friendly. The level of transparency is measured as follows: Compressed and/or non readable file format (where applicable) are at a low level of transparency, Lossless compressed and/or human readable file format(where applicable) are considered at a medium level whereas Uncompressed and/or human readable file format (where applicable) are considered at a high level of transparency.

2.3.4 Stability

Stability of a file format is determined by the format's backward compatibility and its frequency of releases [5]. A file format is backward compatible if it provides all of the functionality of a previous version of the format. Frequency of version/extension releases is another indicator of the stability of a file format. A format with more than one release in the last five years is less stable than a format with one

or fewer releases in the same period. The level of stability is an indication that the development of the format follows a managed release cycle. Resources which are not backward compatible and have a high number of version releases have a low stability level, whereas resources which are backward compatible or have a low number of version releases are considered at a medium level of stability and resources which are both backward compatible and have a low number of version releases are highly stable.

2.3.5 *Interoperability*

Interoperability is the ability of a file format to be accessible on multiple hardware and software platforms [18]. Formats that are supported by a wide range of software or hardware are highly desirable in many situations. This feature also tends to support the long-term sustainability of data by facilitating the possibility of migration of the data from one technical environment to another. Following is the assessment criteria for interoperability: Platform dependent resources are at a low level of interoperability, software interoperable file formats are at a medium level whereas highly interoperable file formats are both software and hardware interoperable.

Scores obtained from each of these factors are aggregated to obtain an overall score. The TAP model considers an aggregated score of 90% and above as a high level of projected preservability and promote such files as recommended file formats, aggregated score of 60% to 90% as a medium level of projected preservability and consider these files as acceptable file formats, and resources below 60% are at the low level of projected preservability and are considered as bit-level file formats. Table 4 shows projected preservability of several file formats.

3. IMPLEMENTATION

Organizations may bundle their preservation strategies based on the preservation level of a resource. There is a lack of agreement on the appropriate number of levels of preservation; the literature contains examples of two [12], three [11], and four [1], to list a few. At the University of Alberta Libraries (UAL), we have resources that we intend to preserve over the long term as well as others that we intend to preserve only over the short and medium term so we have chosen to use three levels of preservation: gold, silver and bronze. Digital resources at the gold level are subject to more rigorous preservation actions than those at the silver or bronze level. The value matrix described in the next paragraph helps to determine the required preservation level of a resource.

3.1 The Value Matrix

The Value Matrix helps to determine the level of preservation for a resource and is based on the three factors mentioned above: type of resource, archival responsibility and projected preservability. Scores obtained from each of these factors are aggregated to obtain an overall score as a guideline to determine the level of preservation appropriate for a resource. UAL suggests an aggregated score of 90% and above to preserve a resource at gold level, 60% to 90% for resources at silver level, and resources below 60% for bronze level. These scores are only used as a guideline; the final decision about the level of preservation for a resource is made by the stewards, curators and technical experts at UAL. Table 3 provides an example of a value matrix.

Preservation Strategies	Gold Plan	Silver Plan	Bronze Plan
Bit Preservation	✓	✓	✓
Core Metadata	✓	✓	✓
Virus Checks	✓	✓	✓
Multiple Copies	✓	✓	✓
Integrity Checks / Checksum	✓	✓	✓
Bagit File Packaging	✓	✓	✓
Unique and Persistent Identifiers	✓	✓	
Normalization	✓	✓	
Characterization and Validation Checks	✓	✓	
Extended Metadata	✓	✓	
Migration	✓		
Full Metadata	✓		
Media Refresh	✓		

Figure 1: Preservation strategies at various levels.

3.1.1 *Gold Level Preservation*

Resources preserved at this level are subject to a rich set of preservation actions for long-term accessibility. Upon ingest, a resource will go through virus checking, fixity checking, file validation, format normalization and archival packaging processes. Gold level resources are archived with full metadata to capture information about the resource, provenance, authenticity, preservation activity, technical environment and rights. To prevent a loss of access to files due to file format obsolescence, all resources at Gold level are subject to a file format migration strategy, which helps to keep the content stored in formats that are readable by the current technology.

3.1.2 *Silver Level Preservation*

Silver level preservation is intended for resources that require medium to long-term preservation but are currently being preserved elsewhere and/or have lower projected preservability. Resources within this plan undergo virus checks, integrity checks, and file format normalization, and include extended metadata. The file format normalization process helps to store resources in UAL recommended archival file formats. Active monitoring is not part of this plan, and it also lacks any migration strategies. Multiple copies help to encounter the problem of media decay and ensure bit-level preservation.

3.1.3 *Bronze Level Preservation*

Resources preserved at this level are subject only to bit-level preservation activities. Under this level, a resource will be subject to virus checks and fixity checking. Only core metadata is archived along with the resource. This is a basic level of preservation which ensures the integrity of each bit over time. Multiple copies of a resource are retained to encounter the perils of media decay and help to replace any corrupted bits with a valid copy. This level of preservation lacks advanced preservation activities like format normalization, format migration, validation checks and full metadata.

UAL uses varying levels of preservation strategies for its gold, silver and bronze resources as shown in Figure 1.

A UAL collection of strength example is the Western Canadiana material held in the Special Collections Library. Much

Table 3: Example of Projected Preservability

File Format	Adoption	Openness	Transparency	Stability	Interoperability	%	PP	Score
xml	2	2	2	2	2	100%	High	3
pdfa	2	2	2	2	2	100%	High	3
rtf	1	0	1	2	2	60%	Medium	2
bmp	1	0	2	0	0	30%	Low	1

Table 4: Example of a Value Matrix

Type of Resource	Archival Responsibility	Projected Preservability	%	Level
5	2	3	100%	Gold
4	2	2	80%	Silver
2	1	1	40%	Bronze

of this material is digitized to the highest possible standards (e.g. jpeg2000, METS/ALTO metadata), and is preserved locally. This type of collection will receive higher scores for all three of the factors considered and therefore could be preserved at the gold level.

UAL's institutional repository contains several collections of locally-created born digital resources, such as photographs and field notes. UAL has less control over file format specifications and so the score for projected preservability could be at acceptable level. The score for archival responsibility remains the same as preservation is local only. This type of collection could be preserved at the silver level.

UAL provides licensed access to a multitude of datasets in support of the research and teaching of faculty and students. Many of these do not fall within our collections of strength, and therefore receive a lower score in terms of type of resource. Because these datasets are created by outside individuals or organizations, file formats vary, with many falling into the 'bit-level' category; projected preservability is therefore lower. Because other institutions (likely including the creator/vendor) also archive these datasets, the score for archival responsibility is lower. As these resources receive an overall lower score hence could be preserved at the bronze level.

4. CONCLUSION

In this paper we have proposed a tiered model for preserving digital content at memory institutions that is built on an assessment which considers three factors: resource type, archival responsibility, and level of projected preservability. This model allows institutions to assess and rank digital resources in terms of preservation needs and then bundle preservation strategies accordingly. We believe the model is simple to apply and flexible enough to be usable by a variety of memory institutions. Although we have described the way in which we have implemented the model at the University of Alberta Libraries, the model does not dictate the method of implementation or the specific preservation strategies to be employed.

5. REFERENCES

[1] N. D. S. Alliance. Ndsa levels of digital preservation: Release candidate one., 2012.

[2] O. I. D. Archive. Digital preservation policies. Technical report, Odum Institute, 2011.

[3] U. D. Archive. Preservation policy, 2011.

[4] A. Bia, R. Munoz, and J. Gomez. Dicomo: the digitization cost model. *International Journal On Digital Libraries*, 11(2):141–153, 2010.

[5] A. Brown. Digital preservation guidance note: Selecting file formats for long-term preservation. the national archives, 2008.

[6] R. Davies, P. Ayris, R. Mcleod, H. Shenton, and P. Wheatley. How much does it cost? the life project – costing models for digital curation and preservation. *Liber Quarterly: The Journal Of European Research Libraries*, 17(1-4):233–241, 2007.

[7] M. Day. Toward distributed infrastructures for digital preservation: The roles of collaboration and trust. *The International Journal of Digital Curation*, 1(3), 2008.

[8] B. Lavoie and L. Dempsey. Thirteen ways of looking at digital preservation. *D-Lib Magazine*, 10(7-8), 2004.

[9] Library and A. Canada. Local digital format registry(ldfr). file format guidelines for preservation and long-term access, 2013.

[10] Y. U. Library. Yale university library policy for the digital preservation. online, 2007.

[11] U. of Minnesota Digital Conservancy. University digital conservancy preservation policy, 2009.

[12] O. C. of University Libraries. Preservation implementation plan., 2011.

[13] U. of Utah J. WIllard Marriott Library. Digital preservation program: Digital preservation policy. Online, 2012.

[14] B. Ooghe and D. Moreels. Analysing selection for digitisation: Current practices and common incentives. In *D-Lib Magazine*, pages 9–10, 2009.

[15] A. Prochaska. Digital special collections: the big picture. *RBM: A Journal of Rare Books, Manuscripts, and Cultural Heritage*, 10(1):13–24, 2009.

[16] J. Rog and C. van Wijk. Evaluating file formats for long-term preservation. national library of the netherlands; the hague, the netherlands, 2008.

[17] K. Skinner and M. Schultz. A guide to distributed digital preservation. Educopia Institute, 2010.

[18] M. Todd. File formats for preservation. dpc technology watch series report 2009. Report, DPC, 2009.

[19] C. Webb. Digital preservation: A many-layered thing: Experience at the national library of australia. In *In Proceedings of The State of Digital Preservation: An International Perspective Conference*, 2002.

Digital Preservation Center of NSLC

Zhenxin Wu
National Science Library, Chinese
Academy of Sciences
33 Beisihuan Xilu, Zhongguancun
Beijing P.R.China ,10019
+86-(10)-82628382
wuzx@mail.las.ac.cn

ABSTRACT
This paper briefly introduces the context and major functions of Digital Preservation Center of National Science Library, Chinese Academy of Sciences (NSLC), and describes its digital preservation system, as well as its preservation services and future work.

Categories and Subject Descriptors
H.3.6 [**INFORMATION STORAGE AND RETRIEVAL**]: Library Automation – Large text archives

General Terms
Management.

Keywords
Digital Preservation Center, Digital Preservation System, Preservation Services.

1. CONTEXT
It is generally acknowledged that digital literature has become the main mode for creating, publishing and disseminating academic information in science and technology fields. This is true in China and globally. The most important function for the staff at the National Science Library, Chinese Academy of Science (NSLC) is to guarantee access to this literature, not only for researchers of the Chinese Academy of Sciences (CAS) who have come to rely on this literature but also with mandate for researchers in the natural sciences and high technology fields across China.

NSLC has been working to build digital infrastructure for digital resources management, in which long-term preservation is regarded as one of key importance, commanding increased attention and investment. A key part of this infrastructure is a digital preservation system (DPS) [1] for preserving digital resources purchased from commercial publishers. Recently NSLC has signed preservation agreements with six publishers and has signed preservation service agreement with three domestic information institutions in the national archive manner. At the time of writing, there are 23,633 electronic journals archived in DPS, from Springer Verlag, Institute of Physics Publishing (IOP), Nature Publishing Group (NPG), BioMed Central, Royal Society of Chemistry (RSC), Chinese VIP STM journals (VIP): more than 28 million articles and 100 million files. The development of processes for preserving eBooks is in progress.

The Institutional Repository Grid (IRGrid) [2] is another component in the digital infrastructure at NSLC, for collecting and storing publications written by researchers of the Chinese Academy of Science. Additional components are being added. Later in 2013, information will be launched about web archiving for the networked resources regarded as important for science and technology and preservation of scientific data.

NSLC has contributing to efforts for long-term preservation over a ten year period, promoting national developments and participating in international meetings, including hosting the iPRES conference in 2004 and 2007. It also now reports its activity on archiving e-journal content to the international Keepers Registry [3] facility.

2. DIGITAL PRESERVATION CENTER
Staff at the National Science Library, Chinese Academy of Science (NSLC) identified four functions [4] for a Digital Preservation Centre (DPC) [5].

(1) Strategy & Planning. It is important at the outset to clarify the objectives for the intended preservation services, defining the scope and selection criteria for archival content, and determining the preservation procedures and processes.

(2) Rights Protection & Management. The interests of stakeholders must be taken into account in order to keep access to resources sustainable. Stakeholders include all parties in the supply chain: libraries as purchasers and their users, publishers, service providers and agents, and the authors/producers of the literature. NSLC implements a comprehensive rights protection and management for its digital preservation.

(3) A Trusted National Archiving System. NSLC wished to provide nationwide preservation services that complied with international standards, referencing international best practices. This is in order to provide entire preservation life-cycle management using scalable technology, one that could be sustainable, reliable and efficient. The infrastructure for preservation is gradually forming for the provision of preservation services nationwide.

(4) Promoting a Cooperative Preservation Network. NSLC has been dedicated to promoting the development of digital preservation nationally through cooperation with major domestic libraries and information institutions. This includes developing and sharing policy and practice, knowledge of standards, and sharing digital preservation services countrywide. An initial step has been a collaborative network for coordinated distribution of multiple secure copies and replacing each other to provide preservation services when necessary. Furthermore, public certification and audits which ensure standardized and transparent cooperation management will be provided within the network.

3. DIGITAL PRESERVATION SERVICES
NSLC has succeeded in its planning and implementation and has established a digital preservation system. This is initially being operated for NSLC and three national organizations with

agreement to preserve the e-journals which they subscribe from Springer, with provision for public access given a trigger event.

The digital preservation system is a dark archive system with have two service platforms:

(1) The archival data management platform. Only accessed by representatives of the national organizations that have agreed to archive their content, this web platform enables auditing of the archived resources, with facility for regular automatic report sent by the DPS. It also allows management of the associated subscription information.

(2) The public access platform. This is intended for the users of the national organizations and for access to subscribed content which may have been triggered according to certain procedures. Similar to the common publishers' service platform, this platform provides browse and search functions, as well as full-text download restrictions based on the subscription information with monitoring of malicious download behavior, providing monthly usage statistics and reports.

Regular audits already have been planned on the schedule and will be carried out by a third party expert group drawn from the National Science and Technology Libraries Group [6] and the Chinese Academic Library and Information System. There is also reporting into The Keepers Registry: for example, search http://thekeepers.org for 'Chinese Journal of Chemical Physics' (1674-0068). In the future, NSLC will provide more services, such as public certification and audit services.

4. DIGITAL PRESERVATION SYSTEM

The IT system department of NSLC is responsible for designing and developing the DPS, as a digital preservation system that was in compliance with the Open Archival Information Systems (OAIS) standard [7]. This includes systematic procedures and policies for the entire lifecycle management. Meanwhile audit management and access control based on preservation agreement have been provided, and system security management and multi-level disaster recovery mechanism have been established, noting the Trustworthy Repositories Audit and Certification (TRAC) standard [8].

The system architecture of the DPS is shown in Figure 1.

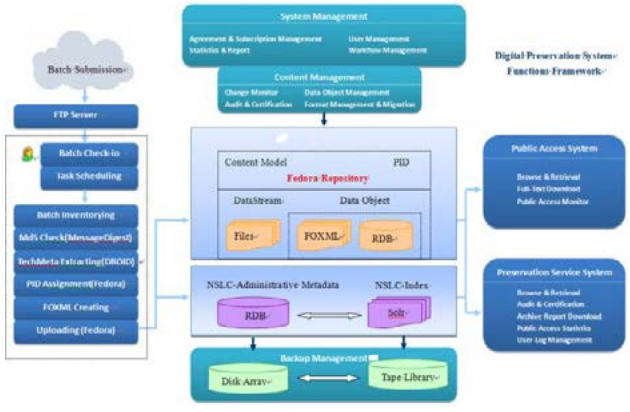


Figure I. Digital preservation system function framework

The core of DPS is the open-source Fedora repository suite [9]. This meant much custom extending and developing had to be done. The DPS can ingest different formats included in the submitted information packages (SIP) and transform these into unified format information packages according to Fedora digital object model, and then store these in the Fedora repository. Utilizing the API provided by Fedora, DPS provides many functions to manage the digital objects, such as fixity checks, version management, etc. With an external relational database, the DPS can store and manage all the metadata which be required for the preservation life-cycle management, from the ingest of the SIP to access of the archived content. With indexing documents, the DPS is easier for browsing, retrieving and statistics. The DPS established at the NSLC is able to provide a complete ingest workflow management and basic preservation management.

4.1 Customizable Ingestion Process

As noted, the DPS can receive and process digital content in a variety of formats, and then generate a unified format for each SIP ingested. Therefore there is a great demand to provide a more flexible workflow mechanism, which can be customized for processing different format SIP. Based on the concept of modular programming, the ingest stage is divided into several modules each with minimum function, such as batch inventory, virus detection, SIPs backup and SIPs transmission, package unzipping, document format validation and technical metadata extraction, metadata verification, standard SIP creation, standard SIP validation, SIPs uploading. The archivist can deploy an appropriate mix by combining different modules to meet the special requirement for a particular SIP.

Function Module Description:

(1) Batch Inventory. Creating and submitting an inventory list of SIPs in order to carry out MD5 checks by using the Java Message Digest [10]. This ensures that the SIPs are unchanged during the transmission process, and later, this list will be use to review against list submitted by the publishers.

(2) Virus Detection. DPS runs Kaspersky [11] on all SIPs for virus scanning.

(3) SIPs Backup and SIPs Transmission. By using the API of JAVA FTPClient [12], DPS replicates the SIPs onto a pre-specified FTP site as a backup, copying to the specified working directory, which is prepared for the next step in the ingest process.

(4) Unzipping Package. DPS runs tar or unzip command to decompress different format data package.

(5) Document Format Validation and Technical Metadata Extraction. Taking into account the efficiency and the quality of the metadata extraction, DPS use Apache PDFBox [13] for PDF format validation and technical metadata extracting, and uses Droid [14] for other format documents.

(6) Metadata Verification. According to the metadata specification agreed with the publishers, DPS verifies the metadata content by using SAX [15].

(7) Standard SIP (FOXML) Creation. DPS adopts the Fedora FOXML [16] as the standard SIP format. It uses SAX to parse the original XML file and extract associated metadata to create a FOXML file, then generates a unique identifier (PID) using the Fedora API-M, establishing the relationships between objects and their data streams. The original XML files, PDF files and other multimedia files are

copied onto the designated directory, as parts of a data object, to be uploaded with the FOXML file.

(8) Standard SIP Validation. Before uploading, DPS once again checks all of the components of the SIPs (including internal and external content).

(9) Uploading SIPs. DPS provides local and remote uploading modes. In the remote mode, DPS uses the Fedora SOAP APIs directly in order to ingest a SIP into Fedora, This keeps flexibility. In the local mode, DPS uses Fedora's underlying function directly without using APIs, which greatly improves the efficiency of uploading.

4.2 Basic Preservation Management

Taking advantage of the features and functions in Fedora, DPS has developed many basic preservation management functions. This use of the Fedora API-A and API-M includes:

(1) Browsing and retrieval of archival data
(2) Multi-level (collection, journal, paper) audit (integrity & fixity)
(3) Archival data maintenance
(4) Tracking changes on data objects
(5) Statistic & Report
(6) Document format management and data migration.

4.3 Agent Execution Mode

It is important to monitor and look for ways to reduce human resource costs and improve efficiency of the system. For this purpose, DPS develop an agent module to execute ingest and other processes deployed by the archivist. The agent module runs automatically in the background which greatly improves the automation level of system. For example, after receiving the data package submitted by publishers, the archivist logs into the DPS to register the receipt of the data and to customize the ingest task. This includes checking the predefined profile (including the designated backup directory and the designated work directory, the selected workflow, etc.) and task scheduling. The Agent monitors task instructions and starts the background data process automatically, the results are sent to the archive administrator by email after the task is completed.

5. FUTURE WORK

There is still much to be done, and to be accorded priority of attention and effort. The following list is put forward for comment and consideration:

(1) Signing preservation agreements with more other publisher, in order therefore to ingest more e-journal content
(2) Serving as the core (node) for domestic long-term preservation community, promoting progress of preservation nationwide
(3) Increasing the types of resource archived: web archiving for important network resources and preservation of scientific data
(4) Increasing preservation service agreements with more resources and more customers
(5) Playing the leading role for the national digital preservation network of China.

6. ACKNOWLEDGMENTS

Thanks are due to all my colleagues who contribute to digital preservation activity at NSLC, especially to Honghu Fu, Yuju Wang and Li Qian from IT department.

7. REFERENCES

NOTE: all URLs successfully accessed June 16, 2013

[1] Digital preservation system (limited access), http://dps.las.ac.cn.

[2] Institutional Repository Grid of CAS, http://www.irgrid.ac.cn/.

[3] The International Keepers Registry , http://thekeepers.org

[4] Xiaolin Zhang, Jiancheng Zheng , Zhenxin Wu, etc. 2012.The Long-term preservation strategy of NSLC (Internal document).

[5] Digital Preservation Centre (DPC) of NSLC, http://dpc.las.ac.cn.

[6] National Science and Technology Library (NSTL), http://www.nstl.gov.cn/.

[7] Open Archival Information Systems (OAIS), http://www.iso.org/iso/home/store/catalogue_tc/catalogue_detail.htm?csnumber=57284.

[8] Trustworthy Repositories Audit and Certification (TRAC), http://www.iso.org/iso/home/store/catalogue_tc/catalogue_detail.htm?csnumber=56510.

[9] Open-source Fedora repository, http://fedora-commons.org/

[10] Java Message Digest (A class of JAVA which provides applications the functionality of a message digest algorithm, such as MD5 or SHA), http://doc.java.sun.com/DocWeb/api/java.security.MessageDigest

[11] Kaspersky (An antivirus and internet security software), www.kaspersky.com/

[12] FTPClient (A tool which encapsulates all the functionality necessary to store and retrieve files from an FTP server), http://commons.apache.org/proper/commons-net/apidocs/org/apache/commons/net/ftp/FTPClient.html.

[13] Apache PDFBox (An open source Java tool for working with PDF documents), http://pdfbox.apache.org/

[14] Droid (Digital Record Object Identification, is an automatic file format identification tool developed by the National Archives, U.K.), http://sourceforge.net/projects/droid/

[15] SAX (An open source Java-only API for XML), http://www.saxproject.org/

[16] FOXML (A simple XML format that directly expresses the Fedora digital object model), http://fedora-commons.org/download/2.0/userdocs/digitalobjects/introFOXML.html.

Enhancing characterisation for digital preservation

Paul Wheatley
University of Leeds
Brotherton Library
Woodhouse Lane
+441133435562
p.r.wheatley@leeds.ac.uk

Gary McGath
Independent Software Developer
42 Heather Court
developer@mcgath.com

Petar Petrov
Creative Pragmatics
Kolonitzgasse 9/11-14
Vienna
petar@creativepragmatics.com

ABSTRACT
Advances in digital preservation software tools have sometimes been slow and or poorly directed. The result has been a lack of tools that meet practitioner needs, and a surplus of tools that have very few users and little practical application. The Jisc funded SPRUCE Project has championed the recording and sharing of practitioner requirements, and the development of solutions to meet those requirements using agile hackathon or mashup style events. This poster will provide a visual summary of requirements identified by practitioners, and will describe four resulting tool developments that significantly advance our digital preservation capability.

Keywords
Digital Preservation, Hackathon, Mashup, User Requirements, Digital Preservation Tools

1. BACKGROUND AND REQUIREMENTS
Practitioners responsible for managing digital data rely on automated software tools [1] to perform many of the key functions that are typical in archival and preservation workflows. Home-grown digital preservation tools have not always developed at a pace or with coverage sufficient to meet practitioners' needs. Steve Knight observed in last year's iPRES opening keynote: "We are still pretty much talking about the same things. Tools like DROID and PRONOM etc. didn't work properly then, and they still don't work properly now" [2]. The problem has not just been that tool development has been lacking, but that the focus and direction of development energies have been poor. Opportunities to build on existing toolsets and incorporate digital preservation requirements have been missed. Even where potentially useful tools have been developed, they have often struggled to find a user base. Examples of duplication and lack of coordination are common[1].

Over the last couple of years the SPRUCE Project [3] has been championing collaborative events that place a strong emphasis on meeting practitioner needs by re-using and enhancing existing software tools. By facilitating the cooperation of both practitioners and software developers, the outcome of tool development has had increased impact and value.

This poster will provide the background to the practitioner requirements and subsequent development (described in sections 2 and 3 below) by outlining the requirements capture process and then highlighting statistics on the number of events at which requirements were gathered (14), the number of practitioners who contributed requirements (100), and the number of organizations which the practitioners represented (70).

[1] For example see "Digital Preservation Cost Modelling: Where did it all go wrong?", which references ~17 different costing models/tools developed to meet very similar aims: http://www.openplanetsfoundation.org/blogs/2012-06-29-digital-preservation-cost-modelling-where-did-it-all-go-wrong

2. WHAT ARE THE PRIORITIES FOR DIGITAL PRESERVATION PRACTITIONERS?
Practitioners were asked to bring their digital preservation challenges to the Open Planets Foundation (OPF) hackathons, AQuA Project mashups and SPRUCE Project mashups that were held over the last couple of years. Further challenges were contributed by the EU funded SCAPE Project. Some constraints were placed on the scope and focus of these challenges, mainly related to the scale of challenges that could realistically be addressed in a 2 or 3 day hackathon. Practitioners were otherwise left to contribute whatever digital preservation challenges they wanted to have addressed.

All of these challenges (and related descriptions of the data on which they are focused, and the solutions developed to solve them) were captured in different locations on the OPF wiki and were then collated on a single wiki page using Confluence tagging functionality [4]. The result is a detailed record of practitioner requirements and current preservation practice.

Five key themes were drawn from the 140+ preservation issues identified by practitioners:

- Quality assurance and repair of damaged or potentially damaged data or metadata
- Appraisal and assessment in order to inform selection, curation and next steps
- Locating preservation worthy data, typically where mixed with other data across shared server space
- Identifying preservation risks in order to inform preservation planning
- A long tail of miscellaneous issues including contextual issues, data capture, embedded objects, and broader issues around value and cost

The overriding focus of these themes is the need to characterize digital data and therefore better understand what it is and what condition it is in. This understanding is typically required before subsequent steps in preservation and curation are undertaken.

This poster will summarize these prioritized practitioner needs, and highlight their relevance for steering future tool development activity.

3. CHARACTERISATION TOOL DEVELOPMENT BASED ON PRACTITIONER NEEDS
Many of the practitioner challenges were tackled as part of the events in which they were raised, with a range of outcomes. Some

resulted in completed tools that were subsequently put into production use at the practitioners' organizations. Some provided proof of concepts or prototypes pointing the direction for future development. Some resulted in unsuccessful approaches, and some remained unsolved.

Analysis of the practitioner needs provided a review point at which to consider next steps for further exploitation of the best work taken on during the hackathon and mashup events, and to consider how the high priority needs could be addressed more effectively. Given the clear need for better characterization, it was decided to host a developer only event which would enable a more concerted effort to update and enhance key digital preservation characterization tools. Further development work could be supported through SPRUCE Awards of up to £5000, which were made available under a funding call for event participants.

A dedicated characterization hackathon was hosted by SPRUCE and the University of Leeds in March 2013 [5]. It was attended by a group of experts including representatives from many of the high profile, home grown digital preservation characterization tools including: JHOVE, JHOVE2, DROID, FIDO, C3PO and FITS. The theme of the event was to coordinate and combine efforts and technology to improve characterization capability.

Four key areas were tackled at the event which are briefly summarized below.

3.1 Solving the PDF Preservation Problem

PDF issues were a recurring theme in previous mashup and hackathon event theme that resulted in a variety of experiments. The majority of these utilized Apache Preflight (or related PDFBox libraries) suggesting this technology had considerable potential. The practitioner challenges also highlighted the inadequacy of existing community solutions. JHOVE for example provides very detailed output for PDFs, but without a clear focus on preservation risks (the main practitioner need) and with data on some risks lacking. Therefore the largest of the four groups at the characterization hackathon wrapped Apache Preflight as a PDF risk analysis tool. Evaluation with large amounts of real data and possible incorporation into key repository technologies to achieve maximum impact for UK Higher and Further Education practitioners (eg. EPrints and DSpace) is being explored at the time of writing.

3.2 Consolidating File Format Identification

The "big 3" file format identification tools, DROID, Tika and File, all have their own file format signatures or "magic" [6], stored in different formats. This data is used to distinguish between different file formats. This leads to the different format identification tools reporting different results for the same file. Each tool has strengths and weaknesses present in its file format magic. Combining the magic would enable a significant improvement in identification coverage and a reduction in inadequate and confusing results for the tool users. Both would be big wins for practitioners. The group made considerable progress in mapping Tika magic to DROID magic. Although not a complete solution, it provided a lot of valuable data for the DROID team to collate and enhance the DROID magic, taking us much closer to a single source for file format magic.

3.3 Wrapping Tika for use in FITS and C3PO

The final two groups looked at addressing the complex picture [7] surrounding the key preservation tools: Apache Tika, FITS and C3PO. All of these tools have considerable potential to deliver effective digital collection assessment via automated characterization, but their current status presents a variety of challenges for end users. FITS for example wraps a number of out of date tools, while C3PO does not offer many extension points.

Two groups of developers at the characterization hackathon focused on incorporating the Apache Tika characterization tool into FITS and C3PO with the aim of making use of the better performance Tika provides and reducing metadata sparsity. Follow up SPRUCE funding awards were granted to address a variety of issues with FITS and C3PO, with the aim of refreshing this toolset. As well as enhancing the functionality and capability of the tools work behind the scenes on the source code and on new documentation has simplified the process for other developers to add support for new tools. This should make future development and support from the community (rather than just the original authors) a more realistic prospect. The OPF will continue to provide coordination, code management, testing and quality assurance to support this process. Further hackathons (such as iPREShack [8]) will provide stimulus for new community sourced developments.

The poster will summarize the tool developments in these four areas, demonstrating how a strongly practitioner led approach can result in well focused tool development and a high impact for the end user.

4. REFERENCES

[1] Digital Preservation Tools http://wiki.opf-labs.org/display/SPR/Digital+Preservation+Tools

[2] Steve Knight quote in: Angevarre, Inge, NCDD Blog, http://www.ncdd.nl/blog/?p=3338

[3] SPRUCE Project, http://wiki.opf-labs.org/display/SPR/Home

[4] Digital Preservation and Data Curation Requirements and Solutions, OPF wiki page, http://wiki.opf-labs.org/display/REQ/Digital+Preservation+and+Data+Curation+Requirements+and+Solutions

[5] SPRUCE Hackathon Leeds: Unified Characterisation, http://wiki.opf-labs.org/display/SPR/SPRUCE+Hackathon+Leeds%2C+Unified+Characterisation

[6] File format magic, Wikipedia http://en.wikipedia.org/wiki/File_format#Magic_number

[7] To fits or not to fits, Petar Petrov, http://www.openplanetsfoundation.org/blogs/2012-07-27-fits-or-not-fits

[8] iPREShack, hackathon at iPRES2013 http://wiki.opf-labs.org/display/SPR/iPREShack+-+SPRUCE%2C+OPF+and+CURATEcamp+hackathon+at+iPRES2013

Query Suggestion for Web Archive Search

Miguel Costa, João Miranda, David Cruz and Daniel Gomes
Foundation for National Scientific Computing
Lisbon, Portugal
{miguel.costa, joao.miranda, david.cruz, daniel.gomes}@fccn.pt

ABSTRACT

Users frequently mistype queries and blame the web archive for poor search results. The addition of a query suggestion functionality in the Portuguese Web Archive had great impact on the perceived quality of the service. In this work, we tested five existing solutions over two datasets. However, existing solutions do not work well, because they rely in predefined lexicons to detect misspellings. We improved the best solutions with a set of rules automatically tuned with an index of archived web collections. The final result can be tested at `http://archive.pt` and the software is publicly available as an open source project.

1. INTRODUCTION

Misspelled queries are common in search engines. Dalianis measured that 10% of web search engine queries were misspelled [1]. Wang et al. counted as misspellings 26% of the total of unique query terms [5]. These numbers explain why most commercial web search engines have a query suggestion module integrated in the user interface. We analyzed a random sample of 1 000 queries of the Portuguese Web Archive (PWA) and detected that 5% were misspelled. This fact was also observed during usability tests, where users were unaware of their mistakes and attributed the poor results to the system's lack of quality [2]. Notice that the PWA returns results even for misspelled queries, because there are documents that contain the same misspelled terms. However, these results are likely not relevant to fulfill the users' information needs.

This work analyzes existing solutions for query suggestion in web archives. As far as we know, this is the first time that such a study was performed and the subject discussed. Our results show that Hunspell[1] optimized with a set of rules provided the best results. We made available the source code of this solution along with a testing dataset of misspellings for evaluation.

This paper is organized as follows. In Section 2, we detail the datasets used in tests. In Section 3, we present the evaluation methodology and the obtained results in Section 4. Section 5 explains how we integrated the chosen solution in the user interface and Section 6 finalizes with the conclusions.

[1] `http://hunspell.sourceforge.net/`

term	misspelling
ameaça	amiaça
coração	corassão
excluir	escluir
higiénico	igienico
manjerico	mangerico
rédea	rédia

Table 1: Example of entries in misspelling datasets.

2. DATASETS

We used two different datasets composed by pairs of <term, misspelling> written in Portuguese. The datasets, named Miranda and Medeiros after their creators, are available at `http://www.linguateca.pt/Repositorio/CorrOrtog/` and contain 394 and 3 890 entries, respectively. Table 1 gives an example of entries in these datasets. At the same site, there are other datasets that could also be used for evaluation.

The Medeiros dataset has a large coverage of typographic and linguistic errors [3]. However, it is 16 years old and was not created having the language used on the web as the main focus. Another dataset was desirable in order for the evaluation to be less prone to errors and overfitting (i.e. fits training data closely, but fails to generalize to unseen test data). Hence, we created the Miranda dataset based on lists of common typos and linguistic errors available on the web, such as `http://ciberduvidas.pt/glossario.php`. This dataset was manually validated by two people.

The variety of Portuguese taken into account was the European Portuguese before the Portuguese Language Orthographic Agreement of 1990. This agreement is an international treaty meant to unify the orthography for the Portuguese language in the countries where it is an official language. Using the official Lince software[2], we found that 98.2% of the entries of the Miranda dataset were compatible with the new norm. Thus, this dataset can be used to evaluate query suggestion algorithms adjusted for a pre or post norm. The results in both cases will be almost identical.

3. METHODOLOGY

Both datasets were split in half, where the first part was used for training the query suggestion algorithms and the second one for testing them. Then, for each entry of the testing part of each dataset, we tested seven algorithms.

[2] `http://www.portaldalinguaportuguesa.org/lince.html`

	Miranda dataset			Medeiros dataset		
	match	**not answered**	**mismatch**	**match**	**not answered**	**mismatch**
Levenstein	4.6%	86.8%	8.6%	4.2%	84.4%	11.4%
Jaro-Winkler	6.1%	70.6%	23.4%	4.3%	61.9%	33.9%
N-gram	1.5%	87.8%	10.7%	2.3%	81.6%	16.0%
Aspell	65.0%	10.7%	24.4%	62.1%	11.6%	26.3%
Hunspell	73.1%	9.6%	17.3%	74.2%	8.7%	17.1%
Aspell+Rules	74.1%	14.7%	11.2%	66.1%	17.6%	16.2%
Hunspell+Rules	77.7%	12.7%	9.6%	76.6%	9.5%	13.9%

Table 2: Results of the query suggesters tested.

These algorithms return a list of suggestions sorted by similarity for each term of a query. This is the usual behavior of spell checking software, which provides several suggestions for the users to choose. However, we followed a web search engine strategy and present only one suggestion in the user interface for not overloading users with too many options. As result, we evaluated as a *match* only when the most similar suggestion provided by the algorithm was equal to the expected term in the dataset. Otherwise, and even if the suggestion was acceptable, we considered it a *mismatch*. Suggestions were *not answered* if the similarity was below a threshold tuned in the training phase.

Let's imagine that for the misspelling *resarcher* the expected suggestion in the dataset is *researcher*. Thus, there is a match if the first suggestion returned by an algorithm is *researcher* or a mismatch if the first suggestion returned is *searcher*.

4. RESULTS

Table 2 presents the obtained results for all tested algorithms over the two datasets. Evaluation measures such as precision (i.e. number of matching suggestions over the total number of suggestions made, $\frac{match}{match+mismatch}$) can be derived from these results.

The three spell checkers available in Lucene 3[3], based on the Levenstein, the Jaro-Winkler and the N-gram distances, yield the lower results as shown in Table 2. For instance, the Levenstein algorithm matched 4.6% of all suggestions in the Miranda dataset and mismatched 8.6%. We tested two other popular solutions: Aspell[4] (version 0.60.3) and Hunspell (version 1.2.9) that greatly improved the results in both datasets. However, the level of mismatch was still high. For instance, Aspell matched 65% in the Miranda dataset and mismatched 24.4%. After we analyzed Aspell and Hunspell more deeply, we applied a set of rules to the suggestions provided by these algorithms by the following order:

1. Suggestions with a difference in length larger than 2 characters when compared to the query term length are ignored. Most of the misspellings only have one or two-character edits (adding, updating or removing). For instance, a suggestion *archer* for the misspelling *resarcher* is ignored.

2. Suggestions split in two (with hyphen or space) are ignored, because they usually mismatch. For instance, a suggestion *res-archer* for the misspelling *resarcher* is ignored.

3. A set of normalizing rules considering the most usual Portuguese misspellings are applied to the query term and its suggestions. Then, a suggestion is returned if it matches the term. The normalizing rules include removing diacritics, adding a prefix *h* (silent letter), and replacing from 3 to 1 char patterns, such as *ssa* by *ça* and *ão* by *am* (same phonetic).

4. Suggestions are discarded if the query term has an index frequency higher than a threshold tuned with the datasets' training part. This frequency is the number of documents of a web archive collection where the term is present. The idea is to ignore suggestions for very used terms, such as names of persons, not contemplated in the dictionary used by the algorithms. For instance, suggestions for the query *Obama* are ignored.

5. A suggestion must have an index frequency n times higher than the index frequency of the query term. The n value was tuned with the datasets' training part. The idea is that the suggestion must occur more times in the collection than the submitted term.

The frequency of the submitted terms and their suggestions were obtained from an index over a collection of 118 million documents archived from 2000 to 2007. Having a large temporal span is important, because the terminology and its use evolves throughout time [4]. Thus, big variations in term frequency are smoothed over the years.

Table 2 shows that these rules increased the *match* percentage in both datasets for Aspell and Hunspell, while significantly reducing the *mismatch*. Hunspell tuned with these rules (Hunspell+Rules), presented the best results and, therefore, was the one integrated in the PWA. For instance, it presented a match of 77.7% and a mismatch of 9.6% for the Miranda dataset. Notice, however, that this algorithm is language dependent due to the rule number 3 applied over the Hunspell suggestions. Still, it seems to work well in English, as shown in Figure 1. Further experiments are needed to confirm this.

We detected that the mismatched suggestions from the optimized Hunspell were mostly caused by the lack of the correct terms in the dictionary. It did not contain names of people nor things, that are commonly searched by users. In the future, this dictionary should be augmented with terms extracted, for instance, from query logs. Another improvement should be considering n-grams of at least two terms, instead of computing the similarity for terms individually.

[3] http://lucene.apache.org/java/3_0_1/api/contrib-spellchecker/org/apache/lucene/search/spell/package-summary.html

[4] http://aspell.net/

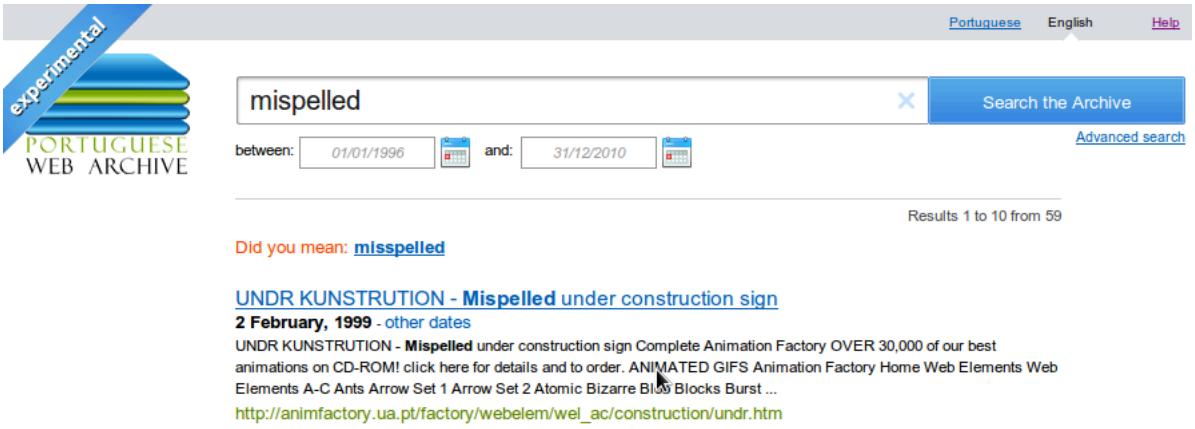


Figure 1: Query suggester integrated in the user interface.

5. INTEGRATION IN THE UI

Figure 1 shows how the query suggestion feature was integrated in the PWA's user interface (UI). This is visible by the *Did you mean* sentence followed by a query suggestion. The suggestion is a link so users can change their query without having to type it again. Our approach was to mimic web search engine interfaces, because users are used to them.

The user interface uses AJAX to make asynchronous calls to the query suggestion service. This enables the searching and query suggesting to be processed in parallel. The searching starts when a user submits a query and the query suggestion request is later triggered after the user's browser starts receiving the results page. Still, the query suggestion response arrives before or soon after the results page has been loaded. Our usability tests conducted on 10 users showed that they did not perceive the asynchronous nature of the query suggester and, thus, they were not distracted by its dynamic behavior [2]. Our tests have also shown that the query suggester is a crucial component for the usability and acceptance of a web archive search service, which led to much fewer negative comments.

6. CONCLUSIONS

Misspelled queries are a common problem in web archives as in web search engines. We tested five existing solutions over two datasets and Hunspell provided the best results. Still, this spell checking software by itself does not achieve a precision high enough to support query suggestion for web archive search. After adding a set of rules to Hunspell, the results were further improved and this is the algorithm that supports the Portuguese Web Archive's query suggester. It can be tested in the production environment at `http://archive.pt`. The software is available as an open source project at `http://code.google.com/p/pwa-technologies/wiki/PwaSpellchecker`.

Many questions remain open that require further research. For instance, should the query suggestion be adjusted to the user's search period of interest? In turn, should the test datasets of misspellings be segmented by time?

7. REFERENCES

[1] H. Dalianis. Evaluating a spelling support in a search engine. *Lecture notes in computer science*, pages 183–190, 2002.

[2] D. Gomes, M. Costa, D. Cruz, J. Miranda, and S. Fontes. Creating a billion-scale searchable web archive. In *Proc. of the 3rd Temporal Web Analytics Workshop*, 2013.

[3] J. Medeiros. Processamento morfológico e correcção ortográfica do português. Master's thesis, Instituto Superior Técnico, Portugal, 1995.

[4] C. Mota. *How to keep up with Language Dynamics: A case-study on Named Entity Recognition*. PhD thesis, Instituto Superior Técnico, May 2009.

[5] P. Wang, M. Berry, and Y. Yang. Mining longitudinal Web queries: Trends and patterns. *American Society for Information Science and Technology*, 54(8):743–758, 2003.

Quality assured image file format migration in large digital object repositories

Using various outcomes of the SCAPE project in the context of library preservation scenarios

Sven Schlarb
Austrian National Library
sven.schlarb@onb.ac.at

Peter Cliff, Peter May, William Palmer
British Library
{peter.cliff, peter.may, william.palmer}@bl.uk

Matthias Hahn
FIZ Karlsruhe
matthias.hahn@fiz-karlsruhe.de

Reinhold Huber-Moerk, Alexander Schindler, Rainer Schmidt
Austrian Institute of Technology GmbH
{reinhold.huber-moerk, alexander.schindler, rainer.schmidt}@ait.ac.at

Johan van der Knijff
National Library of the Netherlands
johan.vanderknijff@kb.nl

ABSTRACT

This article gives an overview on how different components developed by the SCAPE project are intended to be used in composite file format migration workflows; it will explain how the SCAPE platform can be employed to make sure that the workflows can be used to migrate very large image collections and in which way the integration with a digital object repository is intended.

Two institutional image data migration scenarios are used to describe how the composite workflows could be applied in production library environments. The first one is related to the British Newspapers 1620-1900 project at the British Library which produced around 2 million images of newspaper pages in TIFF format. The second is a large digital book collection hosted by the Austrian National Library where the book page images are stored as JPEG2000 image files.

1. INTRODUCTION

Several memory institutions in the SCAPE project, such as the British Library, the National Library of the Netherlands, and the National Library of Austria are using the JPEG2000 image file format for storing images of digital newspapers, books, or other image collections.

In this context the SCAPE project (SCAlable Preservation Environments), partly funded by the European Commission, is doing research and providing solutions that help memory institutions in performing preservation at scale. The project develops an execution platform together with preservation tools and advanced services for preservation planning and watch. Development is driven by institutional requirements and tested in real world institutional environments in order to ensure that the solutions are really applicable on diverse data sets and on a large scale.

This article will give an overview in which ways different components developed by the SCAPE project are intended to be used in composite file format migration workflows, explaining how the SCAPE platform makes these workflows scalable so they can be used to migrate very large image collections. Furthermore, it will discuss the implications that the use of the SCAPE platform has on development and integration of the different components.

We start by explaining the institutional image migration scenario in more detail. We then outline the SCAPE components used in the composite workflows, before presenting the composite workflows themselves. Finally, we conclude the article with a summary and outlook.

2. THE INSTITUTIONAL SCENARIOS

Our first scenario is a real world use case of the British Newspapers 1620-1900 project at the British Library which was funded by the Joint Information Systems Committee (JISC) and produced around 2 million images of newspaper pages in TIFF format [1]. In order to reduce the storage cost of these images, the British Library undertook a migration of the items to the JPEG2000 format prior to ingest into the Digital Library System.

[1] http://www.jisc.ac.uk/media/documents/programmes/digitisation/digitisation_brochure_v2_overview_final.pdf

Our second scenario looks at a large digital book collection hosted by the Austrian National Library where the book page images are stored as JPEG2000 image files and serve as master and access copies at the same time. With this scenario we are showing how an image migration workflow is supposed to work with a digital repository.

For both of these scenarios it is clear that a system capable of migrating millions of images from one format to another is required. Workflows executed on this system must include steps that validate both original and migrated image files and provide assurance that migration was successful and produced equivalent migrated images and valid instances of the new format.

In the following section we describe the SCAPE tools that are used in the composite workflow and which are essential to fulfilling the requirements listed above.

3. PRESERVATION COMPONENTS

As briefly mentioned in the introduction, the SCAPE project is developing new components, extending and improving existing tool implementations and providing means for integration of new tools into the SCAPE preservation platform.

In order to give a complete picture about how software components of different types can be put together in a composite workflow, we make use of two tools developed in the SCAPE project which will be described in more detail in the following sections.

3.1 Jpylyzer

Jpylyzer [8] is a validator tool for the JP2 (JPEG 2000 Part 1) still image format. It was developed to do the verification of whether an encoder produces standard-compliant JP2s, to detect JP2s that are corrupted (e.g. images that are truncated or have missing data), and to extract technical characteristics and metadata.

Although some of the above features are also provided by other software tools, these either provide limited or incomplete validation functionality, partial coverage of JP2's feature set, or produce output that is difficult to interpret. The main philosophy behind Jpylyzer was to create a tool that strictly adheres to the JP2 format specification, is lightweight, simple to use and scalable. The validation procedure includes a verification of the general file structure, tests on the validity of individual header fields, and a number of consistency checks.

3.2 Matchbox

The Matchbox tool was designed for content based image characterization and comparison. It is based on robust detection and invariant description of salient image regions using the Scale Invariant Feature Transform (SIFT) [5]. Categorization of image content uses the Bag of Features (BoF) approach [2] which is inspired by the bag of words approach in information retrieval. In the BoF approach scanned book pages are characterized by compact visual histograms referring to visual words contained in the BoF. The BoF itself is constructed for each collection, i.e. a book scan, using machine learning. Once the BoF is created, image comparison becomes an efficient comparison of histograms. Matchbox also implements detailed image comparison based on the estimation of a geometric transformation between pairs of images followed by the estimation of a perceptual measure of Structural Similarity (SSIM) [9].

4. EXECUTION PLATFORM

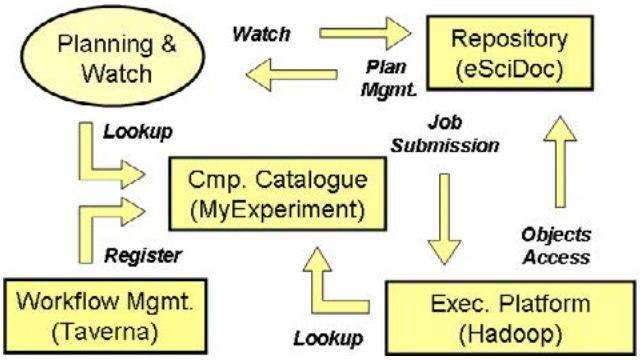


Figure 1: Components and services of the SCAPE Preservation Platform. The available software components provide support for workflow design and description, registration and lookup of preservation components, scalable storage and execution, and digital object management and efficient access. Integration with the SCAPE Preservation Planning and Watch components is supported through the Component Catalogue Lookup API and the Repository Plan Management and Watch APIs.

The SCAPE Preservation Platform [6] provides an infrastructure that targets the scalability of preservation environments in terms of computation and storage. The goal is to enhance the scalability of storage capacity and computational throughput of digital object management systems based on varying the number of computation nodes available in the system. A platform instance is based on existing, mature software components like Apache Hadoop[2], the Taverna Workflow Management Suite[3], and the Fedora Digital Asset Management System[4]. The platform implements a set of additional services on top of these software components to specifically support scalability and integration with digital preservation processes as well as to integrate with other SCAPE components, such as the SCAPE preservation watch system, SCOUT [1]. Figure 1 provides an overview of the main software components of the SCAPE preservation platform and shows their interactions.

A key challenge of the platform is the development of methodologies to integrate preservation tools with its parallel execution environment. The automated deployment of preservation tools such as Jpylyzer, described in section 3.1, is based on software packages like those maintained by the Open Planets Foundation [5] and a Linux based software package management system (presently based on Debian). Complex software environments like pre-configured platform nodes can be deployed on virtualized hardware using virtual machine images[7]. The platform provides support for migrating existing and sequential preservation workflows and applications to the parallel environment covering different aspects like data decompostition, tool handling, workflow support, or repository interaction. However, the strategy

[2]http://hadoop.apache.org
[3]http://taverna.org.uk
[4]http://http://www.fedora-commons.org
[5]http://deb.openplanetsfoundation.org

used to parallelize an individual workflow depends on the use case it implements and may be selected on a case-by-case basis. Section 4.2 discusses basic parallelization approaches with respect to the example workflow discussed in this paper. A flexible mechanism for the integration of existing digital repository systems is provided by the SCAPE Data Connector API. This generic interface supports the efficient exchange of data sets between the execution platform and digital object management systems, as described below.

4.1 Digital Object Repository

The SCAPE platform provides a Digital Object Repository to allow storage and management of digital objects. The repository offers several APIs to integrate with the SCAPE platform and other SCAPE components like Planning and Watch. Preservation actions running on the execution environment are able to interact with the repository via a RESTful service API. This Data Connector API allows ingest, retrieval, update and query of a repository's content.

A Digital Object Model has been defined to allow different SCAPE components to exchange data in a standardized way. This model is based on METS[6] as a container format, along with other metadata formats like Dublin Core[7], Marc 21[8], PREMIS[9] and other technical, administrative and rights metadata. The data we are focusing on is already provided in a METS format and can be ingested into the repository via the SCAPE Loader Application, a Java-based client application supporting different input source options. (local or distributed file system). Its intended use is for ingesting a large amount of digital objects (represented as METS) into the repository using the REST endpoint defined by the Data Connector API. It monitors and logs the ingest process, e.g. retrieves the life-cycle status of each digital object of the repository.

4.2 Scalable Processing

The SCAPE preservation platform utilizes the Apache Hadoop framework as the underlying system for performing data-intensive computations and consequently relies on MapReduce [3] as the parallel programming model. In SCAPE, preservation scenarios are typically developed as sequential workflows using desktop tools like the Taverna workbench. Such conceptual workflows, which will be explained in more detail in section 5, define the general logic of a preservation scenario and must be migrated to the parallel environment before they can be executed on the SCAPE preservation platform at scale.

Depending on their complexity, preservation workflows (or activities within a workflow) can be turned automatically into a parallel application that runs on the platform to a certain degree. An example is the execution of preservation tools against large volumes of files which can be performed on the platform using a generic MapReduce tool wrapper. The SCAPE tool specification language supports users in selecting a particular tool and parameter configuration used during the execution. SCAPE has also developed a model allowing a workflow designer to describe preservation activities following a defined component specification and register them to the SCAPE Component Catalogue (c.f. figure 1).

[6]http://www.loc.gov/standards/mets/
[7]http://dublincore.org/
[8]http://www.loc.gov/marc/bibliographic/
[9]http://www.loc.gov/standards/premis/

The platform makes use of this approach to discover runtime dependencies of workflows, like dependencies on preinstalled software packages, which must be resolved prior to workflow execution.

However, as discussed in this paper, it is typically required to migrate more complex workflows involving different activities, data flows, and decision logic to the platform environment. A simplistic approach is to instantiate and concurrently execute multiple instances of the sequential workflow on a range of cluster nodes. This strategy however comes with a number of restrictions as compared to an approach where the workflow language is fully translated into a native MapReduce program, a strategy which is also evaluated in the context of SCAPE.

5. WORKFLOWS

As already mentioned, Taverna [4] is used in the SCAPE project to build composite workflows using the components described in section 3.

Having created a single-threaded sequential Taverna workflow, as noted in the platform section 4, it is necessary to translate this into a suitable MapReduce program for execution on the SCAPE Platform. Performing actions like file migration using Hadoop is achieved by using one or more map jobs (made up of many map tasks) across a number of processing machines and few (if any) reduce jobs.

5.1 Taverna integration

The workflow in Figure 2 shows the steps required to migrate a TIFF to a JP2 and quality assure the results. It was designed to address the requirements of the British Library's TIFF to JP2 migration scenario. Input to this workflow is a list of TIFF files and the output is the migrated JP2s and a report giving details of the migration and quality assurance stages. The workflow consists of both sequential and parallel layers. For example, once the TIFF to JP2 migration completes (HadoopMigrate) then metadata extraction, feature extraction using Matchbox and profile validation using Jpylyzer can all operate on that JP2 at the same time. Similarly, while TIFF to JP2 migration is taking place, the workflow can also be extracting features from the TIFFs using Matchbox ready for comparison with the features extracted later from the JP2.

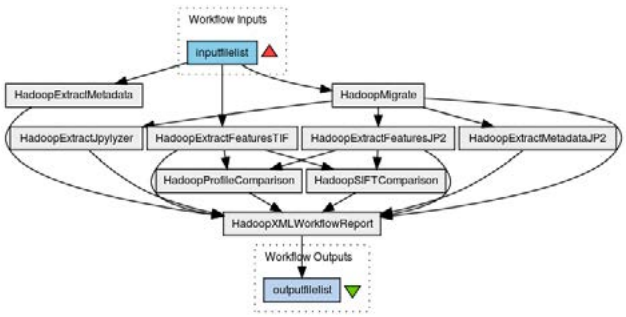


Figure 2: SCAPE Platform migration TIFF to JP2; http://www.myexperiment.org/workflows/3400

When translating a workflow like this we need to decide what each map task should do, and there are different options discussed in the following.

5.1.1 Vertically aligned workflow

Orienting the workflow with inputs at the top and output at the bottom, one option is to slice the problem vertically and execute the workflow from top to bottom for every input file. Here each map task calls the Taverna command line with a single input file and the workflow definition. Taverna is responsible for the order of execution within the workflow and runs steps in parallel, where possible, according to the workflow graph.

This vertical slicing has a number of advantages. Taverna preservation workflows that work well on single machines can easily be scaled using the SCAPE Platform. Workflow designers do not need knowledge of Hadoop and workflows can be re-used. This is the idea behind SCAPE components. We can also make use of Hadoop's robust design: should the workflow fail, that map task fails; Hadoop will handle retrying the map task and reporting the failure. Many workflows will create intermediate files on the processing data node. Doing all the work on a single data node avoids moving these files across the Hadoop cluster and managing their locations. Finally, Hadoop requires no knowledge of Taverna, and (unless using HDFS) the workflow does not need any knowledge of Hadoop.

5.1.2 Horizontally aligned workflow

Another option is to slice the problem horizontally and execute each layer of the workflow as a chain of map tasks. For the workflow presented in Figure 2 the TIFF to JP2 migration is performed over all files, one map task per migration. At the same time a second set of map tasks can be extracting the features and metadata of the TIFFs. Once complete another set of map tasks extract features from the JP2s and so on. It is clear that something is needed to manage this execution and for this we can use Taverna. However, this approach requires that the sequential workflow be re-written with knowledge of Hadoop.

5.1.3 Translation to MapReduce

A final option would be to translate the Taverna workflow to one or more native Hadoop jobs, using Taverna to design the workflow but not using it during execution. This strips away a layer of complexity.

5.2 Digital objects repository integration

The JPEG2000 to TIFF migration scenario using the digital book collection of the Austrian National Library provides a production environment for testing the large scale applicability for the digital objects repository integration.

In this scenario, digital book objects are ingested using SCAPE's Loader Application described in section 4. First, METS containers as the submission information packages (SIPs) according to the OAIS reference model, aggregates the digital book and book page entities (each book page consisting of an image, full text, and full HTML layout representation) with references to the physical files on the file server.

The goal is to find a performant way of doing ingest, migration, and finally adding a new representation to existing digital objects using the SCAPE Platform. Towards the end of the SCAPE project, an evaluation will be made of overall system and component level performance indicators.

6. CONCLUSIONS

In this article we have presented several core outcomes of the SCAPE project along with preservation scenarios that give a better idea of how they can be used in an institutional context. We have also shown how tools can be used in workflows combining characterisation, migration, and quality assurance tasks.

According to the SCAPE project's mission to provide solutions that work on a large scale, we have discussed approaches to transform conceptual workflows into workflows which can be executed on the SCAPE platform and integrated with a digital object repository.

The development of these workflows will be pursued further this year; towards the end of the project, evaluations will give more insight into performance, runtime stability and organisational fit of the solutions presented in this article.

7. ACKNOWLEDGMENTS

This work was partially supported by the SCAPE project. The SCAPE project is co-funded by the European Union under FP7 ICT-2009.4.1 (Grant Agreement number 270137).

8. REFERENCES

[1] C. Becker, K. Duretec, P. Petrov, L. Faria, M. Ferreira, and J. C. Ramalho. Preservation watch: What to monitor and how. In *Proc. of the Ninth Int. Conf. on Preservation of Digital Objects (iPres12)*, Toronto, Canada, October 2012.

[2] G. Csurka, C. R. Dance, L. Fan, J. Willamowski, and C. Bray. Visual categorization with bags of keypoints. In *Workshop on Statistical Learning in Computer Vision, ECCV 2004*, pages 1–22, 2004.

[3] J. Dean and S. Ghemawat. Mapreduce: simplified data processing on large clusters. *Commun. ACM*, 51:107–113, January 2008.

[4] D. Hull, K. Wolstencroft, R. Stevens, C. Goble, M. R. Pocock, P. Li, and T. Oinn. Taverna: a tool for building and running workflows of services. *Nucleic Acids Res*, pages 729–732, 2006.

[5] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *Int. J. of Comput. Vision*, 60(2):91–110, 2004.

[6] R. Schmidt. An architectural overview of the scape preservation platform. In *Proc. of the Ninth Int. Conf. on Preservation of Digital Objects (iPres12)*, Toronto, Canada, October 2012.

[7] R. Schmidt, D. Tarrant, R. Castro, M. Ferraira, and H. Silva. Guidelines for deploying preservation tools and environments. Technical report, SCAPE Project Deliverable, March 2012.

[8] D. Tarrant and J. V. D. Knijff. Jpylyzer: Analysing jp2000 files with a community supported tool. -, October 2012.

[9] Z. Wang, A. Bovik, H. Sheikh, and E. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600 –612, April 2004.

Automating the Preservation of Electronic Theses and Dissertations with Archivematica

Mark Jordan
Simon Fraser University
8888 University Drive
Burnaby, British Columbia, Canada
1-778-782-5753
mjordan@sfu.ca

ABSTRACT
This poster describes the tools, services, and workflows that Simon Fraser University is using to automate the movement of its ETDs (Electronic Theses and Dissertations) from its user-facing Thesis Registration System to the Archivematica digital preservation platform. The poster also describes Simon Fraser University's plans to expand its digital preservation services using Archivematica, including integration of LOCKSS as a distributed storage network for content managed by Archivematica.

Categories and Subject Descriptors
H.3.4 [**Information Storage and Retrieval**]: Systems and Software – *Distributed systems* and H.3.7 [**Information Storage and Retrieval**]: Digital Libraries – *Standards*

General Terms
Management, Standardization

Keywords
Case studies, Digital Preservation, ETDs, workflows, automation, microservices, OAIS, Drupal, Archivematica, LOCKSS

1. INTRODUCTION
Simon Fraser University (SFU) has been accepting theses, dissertations, and graduate project reports from students in digital form since 2004. In late 2012, the Library initiated a set of microservices to transfer electronic theses and dissertations (ETDs) theses from its Theses Registration System (TRS)[1] to its institutional repository, Summit,[2] without human intervention apart from sign off by Library staff that the thesis has become ready for publication. Shortly after the initiation of that automated workflow, the Library started moving theses from the TRS into the Archivematica[3] digital preservation platform, a process which is also fully automated.

This poster describes the rationale for automating the ingestion of ETDs into Archivematica, the various tools and services that are used in this automation, and how they work together. It also describes areas of active development the SFU Library is pursuing to expand this set of digital preservation services.

2. GOALS AND GUIDING PRINCIPLES
Theses and dissertations are one of the most important types of scholarly works created by universities. Even though copies of ETDs are frequently distributed in commercial services such as Proquest Dissertation Publishing[4] or in national aggregations such as Theses Canada,[5] many educational institutions that produce ETDs take on the responsibility for long-term preservation of these works. However, this commitment will require considerable resources over time.

Simon Fraser University has decided to act on this responsibility but to do so with the goal of reducing costs as much as possible. Many of the costs associated with digital preservation are difficult to predict,[6] but one aspect of this activity in which it is relatively easy to minimize costs is human labor. To that end, SFU is striving to automate as many aspects of the ETD lifecycle as completely as possible. Three guiding principles led to the development of a set of services and processes to achieve this goal.

First, the preservation of ETDs should adhere to proven, robust, standards-based digital preservation practices such as compliance with the OAIS Reference Model,[7] use of PREMIS[8] preservation metadata, use of the BagIt[9] content packaging format, and support for standard descriptive metadata such as Dublin Core Terms.

Second, any processes involved in the preservation of ETDs that can be automated should be. Human intervention will be required at certain points in preservation workflows, but the amount of human intervention and localized decision making should be reduced to a practical minimum.

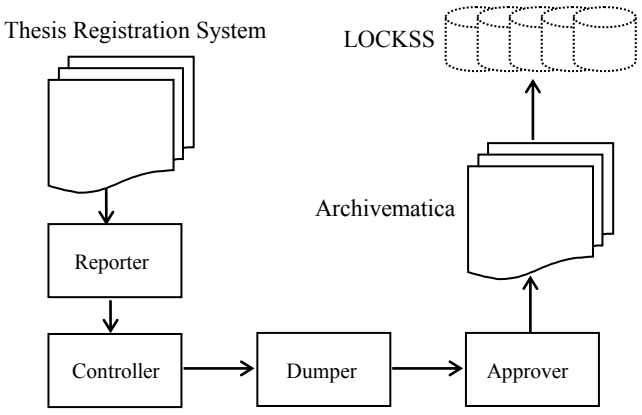
Third, any specific service or tool used in these processes should be easily replaceable. Over the long term, tools considered best in class will invariably change. It is important that any new tool that improves a process, or performs the same process at less cost, should replace the existing tool, as long as doing so is not overly disruptive to other processes that depend on the exiting tool. In addition, the ability to replace services and tools facilitates easy adaptation of the remaining components to other digital preservation processes.

These three principles informed the development of the preservation architecture described below.

It is important to note that an ETD is not only a simple textual document. Many ETDs have raw or application-specific data, multimedia content, or additional textual documents associated with them. This additional content is commonly referred to as "supplemental" content or files. In addition, an ETD will typically have at least one metadata description identifying its title, date of completion, subject content, and so on, usually expressed in the ETD-MS[10] element set. The preservation of an ETD is therefore not as simple as making sure that the thesis document is stored in a single PDF file. Long-term preservation of ETDs must take all of these types of content into account. [11]

3. ARCHITECTURE

Simon Fraser University's ETD preservation architecture is comprised of three main components, 1) its Thesis Registration System, 2) a set of microservices, and 3) the Archivematica digital preservation platform. A fourth component, a Private LOCKSS Network, is currently under development. The following is a visual overview of this architecture:



3.1 Thesis Registration System

The Thesis Registration System enables students to register their thesis and upload any associated files in what is referred to as a "submission." Once the student has completed a submission, Library staff audit the thesis before they approve it for publication in the University's institutional repository. This process involves verifying that the thesis adheres to publication standards set by the University, that any documentation such as ethics review approval has been obtained, and that all licenses for publication have been accepted by the student.

When all audit requirements have been met, Library staff record this decision within the submission record for a thesis by simply checking a box titled "Ready for publication." This attribute of the thesis submission is then used in a query, run nightly, to identify all submissions that have been approved for publication during the previous day.

The Thesis Registration System is built using Drupal,[12] an open-source Content Management Framework. Drupal manages user accounts and permissions, provides mechanisms for structuring the thesis submission, and handles the various types of files the student must upload. A custom Drupal module, developed by SFU Library staff, manages the workflows involved in auditing the submission, and sends out email messages to the student when the audit staff perform specific actions or make specific decisions. Each submission is instantiated within the Thesis Submission System as a "node," the basic content structure within Drupal.

3.2 Microservices

Moving the ETD content out of the Thesis Registration System and into Archivematica is accomplished using a small series of microservices. Each microservice is a shell or PHP script that performs one task or one group of related tasks.

The first microservice (called the "reporter") queries the Thesis Registration System for all submissions that have been approved during the previous day. This is the script that performs the query described in section 3.1, above. The script writes the Drupal node IDs (which serve as the unique identifier of each thesis submission in the Thesis Registration System) to a data file with the current date encoded in its filename.

The second microservice (the "controller") wraps two task-specific scripts; in other words, it runs each of the two scripts from within itself. This approach allows for robust handling of errors in each script, and also allows for easy cleanup of temporary files created by the wrapped scripts. The controller is scheduled to run each day after the reporter microservice runs, and uses the data file created by the reporter as its input. In effect, the controller loops through all of the submission node IDs in the current day's data file and runs the two wrapped microservices, the "dumper" and the "approver," on the submission corresponding to each node ID.

The "dumper" microservice takes a submission node ID as a parameter, queries the Thesis Registration System for the corresponding submission node, and creates Dublin Core and ETD-MS descriptive metadata files for the thesis using information in the submission record. In addition, the dumper microservice determines what files are associated with the thesis (the thesis PDF, any supplemental files, and specific licenses and other administrative documents) and writes those out to disk as well. Finally, the dumper ensures that all of the files are arranged in a subdirectory structure compliant with Archivematica's "transfer" package format (described in the next section) and creates a Bag containing the all the submission's files.

The final microservice is the "approver," which copies the Bag created by the dumper to the Archivematica server and, after confirming that the Bag has been copied successfully, issues an HTTP request to Archivematica's transfer approval API (also described in the next section).

3.3 Archivematica

Archivematica is an open-source digital preservation platform. It normalizes files into preservation-friendly formats using what it calls "format policies",[13] and stores content in OAIS-compliant Archival Information Packages (AIPs). Archivematica integrates a number of open-source tools such as FITS,[14] OpenOffice,[15] FFmpeg,[16] and Clam Antivirus[17] using its own internal microservices framework, and it employs open, standardized formats such as METS,[18] PREMIS, and BagIt to ensure long-term, standards-based management and access to the content and metadata stored in the AIPs it produces.

Content is ingested into Archivematica as a "transfer," which contains the files to be preserved, metadata describing those files, "submission documentation" (licenses and other administrative documents), and, optionally, a "processing configuration" file. The transfer structures the content in preparation for repackaging into an OAIS Submission Information Package (SIP) and then, into an Archival Information Package (AIP) for long-term management. If the content is to be made available to a given user community, Archivematica allows the creation of Dissemination Information Packages (DIPs) for that purpose.

Archivematica's user interface breaks down the workflow for processing a given set of files from transfer to SIP to AIP to DIP into a series of structured tasks, most of which are instantiated internally as microservices. Within each group of tasks, a human operator must make a number of decisions, such as whether to normalize the incoming files for preservation, access (or both),

whether to approve the results of the normalization or not, whether to apply additional descriptive metadata to the transfer, and where to store the AIP. How specific file types are normalized is determined by the format policies; for example, the format policy for audio files might dictate that they should be normalized into WAV format for preservation and MP3 format for end user access.

Workflow tasks can be automated using a processing configuration file, which encodes in a machine-readable format each of the decisions that a human operator would make if he or she were manually processing a transfer. The ability to automate workflow decisions is useful if Archivematica is to process large quantities of similar transfers in batches, or if local policy dictates that a given workflow decision should always be made.

For the processing of Simon Fraser University's ETDs, the processing configuration file specifies that the files should be normalized for preservation only (since we are not asking Archivematica to generate Dissemination Information Packages), which format identification tool Archivematica should use, and where to store the AIP.

The processing configuration file only removes the need for a human operator after a transfer package has been ingested into Archivematica. To automate the ingestion itself, Archivematica provides a REST API[19] for approval of transfers. Since the API uses REST, it is possible to interact with this API from within a script running on a different server (in this case, the "approver" microservice running on server hosting the Thesis Submission System).

It is the combination of this REST API and the processing configuration file that allows for the complete automation of moving content from a source such as SFU's Thesis Registration System into Archivematica, then through Archivematica's digital preservation microservices to produce an OAIS-compliant Archival Information Package. In the case of SFU's architecture for preserving ETDs, this process is instantiated in the dumper and approver microservices described earlier, which combined, hand over the ETD content to Archivematica's internal microservices as defined by the processing configuration file.

3.4 Long-term management of the ETDs

Over time, Archival Information Packages can be retrieved and re-ingested into Archivematica as SIPs (Submission Information Packages) when the content needs to be updated or migrated to new formats. The need to update an ETD after its publication is rare but not unheard of, and SFU's Faculty of Graduate Studies has a policy in place for that situation.

Archivematica supports the authenticity of content it preserves by storing all original documents that are included in a transfer in addition to any normalized versions created by its microservices (or by normalization external to Archivematica). It also generates and stores checksums for all files to allow auditing and verification of bit-level integrity over time. Finally, in SFU's implementation, all license agreements signed by the author of the ETD are preserved in the same Archival Information Package as the ETD document and supplemental files, complete with checksums.

3.5 Public access to the ETDs

The version of the ETD content that is transformed by Archivematica into an OAIS Archival Information Package is not intended to be accessed by end users. In fact, the AIP contains licenses and other sensitive information that should not be exposed to end users.

In SFU's implementation, the ETD and its associated metadata are transferred directly from the Thesis Registration System to the University's institutional repository, Summit, for public access. This transfer is automated and happens at the same time as the transfer of the ETD from the Thesis Registration System to Archivematica. In effect, the two processes are run in parallel. Once in the institutional repository, end users access the theses through a variety of discovery tools such as the Library's unified discovery layer and the search and browse capabilities of Summit itself.

Archivematica is capable of creating an OAIS Dissemination Information Package (DIP) and transferring the DIP to a variety of public-access content management systems and repository platforms, including AtoM, CONTENTdm, and DSpace. SFU's implementation does not use this feature because a process to move ETDs from the Thesis Registration System to Summit was already in operation when the Library began using Archivematica. It would be possible to create new Archivematica microservices to produce a DIP for SFU's Summit, but the Library has chosen an alternative approach to integrating Archivematica and its institutional repository, described in section 4.3, below.

4. DEVELOPMENT ROADMAP

The SFU Library is actively working to expand the integration of its current ETD preservation services with several other tools.

4.1 LOCKSS integration

Work is under way to allow Archivematica to store its AIPs in a Private LOCKSS Network (PLN).[20] This development will enable the automated movement of AIPs into a holding queue, from which LOCKSS will harvest them and ingest them into the PLN. Storing the AIPs in a Private LOCKSS Network will ensure that identical copies are managed in a geographically distributed, secure fashion. SFU Library and a group of partner institutions are working closely with the developers of Archivematica to ensure that this work is compliant with a new Storage API that is being developed for Archivematica. This API will allow it to use a variety of storage platforms for AIPs it generates.

4.2 Academic review of theses using Open Journal Systems

Although not directly related to preservation of ETDs, the Faculty of Graduate Studies at SFU is planning to use Open Journal Systems (OJS)[20] for the academic review of theses. Open Journal Systems provides a toolset for manuscript submission, peer review, and editorial workflow for journal articles that is easily adaptable to the review of theses by academic adjudication committees. SFU Library will be working closely with the Faculty to ensure that ETDs will move from OJS to the Library's Thesis Submission System seamlessly, and from there, through the digital preservation architecture described in this poster.

4.3 Automating preservation of content in SFU's institutional repository

The tools and workflows described in this poster can also be applied to automating the preservation of content submitted to SFU's institutional repository, Summit. Work is under way to

implement such a process. Essentially, all that is required is to modify the dumper microservice to convert non-ETD items in the institutional repository into Archivematica transfer packages. "Non-ETD items" in SFU's repository include journal and book chapter preprints, conference papers, reports, and other works submitted directly by end users and by Library staff as a service to the University community. Automating the movement of content from Summit to Archivematica will provide robust digital preservation services lacking from many institutional repositories.

The ability to replace one component of SFU's digital preservation architecture (the Thesis Registration System) with another (the institutional repository) and make only minor modifications to a single microservice (the dumper) illustrates an important guiding principle of the architecture: "any specific service or tool used in these processes should be easily replaceable." This pattern can also be applied to other sources of content the SFU Library needs to preserve, such as locally digitized manuscript collections and newspapers, research data sets, and archived websites.

5. REFERENCES

[1] Simon Fraser University's Thesis Registration System. https://theses.lib.sfu.ca

[2] Summit, Simon Fraser University's Institutional Repository. http://summit.sfu.ca

[3] Archivematica. https://archivematica.org

[4] Proquest Dissertation Publishing. http://www.proquest.com/en-US/products/dissertations/

[5] Theses Canada. http://www.collectionscanada.gc.ca/thesescanada/index-e.html

[6] Wheatley, Paul. 2012. *Digital Preservation Cost Modelling: Where did it all go wrong?* Blog post. http://openplanetsfoundation.org/blogs/2012-06-29-digital-preservation-cost-modelling-where-did-it-all-go-wrong

[7] Reference Model For An Open Archival Information System (OAIS). http://public.ccsds.org/publications/archive/650x0m2.pdf

[8] PREMIS Data Dictionary for Preservation Metadata. http://www.loc.gov/standards/premis/

[9] The BagIt File Packaging Format. http://tools.ietf.org/html/draft-kunze-bagit-09

[10] ETD-MS: an Interoperability Metadata Standard for Electronic Theses and Dissertations. http://www.ndltd.org/standards/metadata/

[11] Shreeves, Sarah L. 2013. *Supplemental Files in Electronic Theses and Dissertations: Implications for Policy and Practice.* Poster presented at the 8th International Digital Curation Conference, Amsterdam, Netherlands, January 14-17, 2013. http://hdl.handle.net/2142/35314

[12] Drupal. http://drupal.org/

[13] Archivematica Format Policies. https://www.archivematica.org/wiki/Media_type_preservation_plans

[14] File Information Tool Set (FITS). http://code.google.com/p/fits/

[15] OpenOffice. http://www.openoffice.org/

[16] FFmpeg. http://www.ffmpeg.org/

[17] ClamAV. http://www.clamav.net/lang/en/

[18] Metadata Encoding and Transmission Standard (METS). http://www.loc.gov/standards/mets/

[19] Approving a transfer. https://www.archivematica.org/wiki/Administrator_manual_0.10#Approving_a_transfer

[20] Lots of Copies Keep Stuff Safe (LOCKSS). http://www.lockss.org/

[21] Open Journal Systems (OJS). http://pkp.sfu.ca/ojs

Diverse approaches to blog preservation: a comparative study

Richard M. Davis
University of London
Computer Centre
Senate House, Malet Street
London WC1E 7HU
+44 20 7692 1350
richard.davis@london.ac.uk

Edward Pinsent
University of London
Computer Centre
Senate House, Malet Street
London WC1E 7HU
+44 20 7692 1345
edward.pinsent@london.ac.uk

Silvia Arango-Docio
University of London
Computer Centre
Senate House, Malet Street
London WC1E 7HU
+44 20 7692 1343
silvia.arango-docio@london.ac.uk

ABSTRACT

This poster presents highlights of a comparative study of three distinct approaches to preserving the content of blogs, to consider the relative benefits of each approach in meeting the requirements for blog preservation, in different contexts. Assessment criteria are drawn from key publications and frameworks on digital preservation as well as practical considerations derived from the authors' experience as users and designers of digital archiving tools and systems.

Categories and Subject Descriptors

H.4 [Information Systems Applications]: Miscellaneous; D.2.8 [Software Engineering]: Metrics / complexity measures, performance measures

General Terms

Management, Performance, Design, Reliability, Human Factors, Standardization, Theory.

Keywords

digital preservation, digital curation, designated community, authenticity, intellectual entity, archive, web archive, blog, weblog

1. INTRODUCTION

The importance of blogs as a distinct class of Web resource has received considerable attention in recent years, notably at iPRES (Pennock and Davis, 2009 [1]; Kim and Ross, 2011 [2]; Stepanyan et al, 2012 [3]). The need to capture this dynamic, cumulative content for future access has been recognised by several institutions and projects and a variety of tools and approaches have emerged.

This poster will present, in graphic form, a summary of key results of interest from a comparative analysis of three distinct approaches to blog-archiving, each of which differs significantly in its methodology, strategy, and delivered outcomes.

The study is based on key criteria derived from study of a wide range of established frameworks in digital preservation, including:

- Reference Model for an Open Archival Information System (OAIS) [4]
- Preservation Metadata Implementation Standard (PREMIS) [5]
- Metadata Encoding and Transmission Standard (METS) [6]
- Trustworthy Repositories Audit & Certification (TRAC) [7]
- Digital Repository Audit Method Based On Risk Assessment (DRAMBORA) [8]

The study compares the relative strengths of three types of digital archive/repository in the context of blog preservation: one created specifically for blogs; another designed for institutional publications; and a third designed for Web content.

The study identifies a number of indicators for success in web-archiving, and a select range of metrics to the effectiveness of each approach against established criteria, derived from the authors' experience and review of literature on best practice for web archiving projects. The study will be completed during June 2013 and the highlights are presented in the accompanying poster.

2. THREE APPROACHES TO BLOG ARCHIVING

1. The BlogForever project, funded by the European Union, has developed an integrated platform, comprising a harvesting methodology and associated content management system, for creation, management and preservation of blog collections.
2. The London School of Economics (LSE) preserves its academic blogs by creating and depositing PDF renditions of blog posts into an existing Institutional Repository.
3. The UK Web Archive, operated by the British Library, which collects and preserves blog content from the UK Blogosphere. This collection represents a cross section of UK Web logs containing a wealth of material which will be of value to researchers now and in the future.

3. ASSESSMENT AND SELECTION CRITERIA

The assessment criteria are derived from definitions and understanding of digital preservation as expressed in the following standards, projects and reports.

1. Long Term Preservation (OAIS): does the repository offer sufficient control of the content to ensure long-term preservation?
2. Designated Community (OAIS): does the repository identify a Designated Community who should be able to understand the information provided; and is the content independently understandable and available to the Designated Community?
3. Preservation metadata (PREMIS): does the repository support the viability, renderability, understandability, authenticity, and identity of digital objects in a preservation context?
4. Metadata encoding and transmission (METS): is there metadata necessary for both the management of digital objects within a repository and exchange of such objects between repositories (or between repositories and users)?
5. Long-term Access (TDR and TRAC): can the repository provide reliable, long-term access to managed digital resources to its designated community?
6. Digital curation risks (DRAMBORA): does the repository demonstrate it effectively and efficiently manages the risks associated with the process of curating digital materials?
7. Completeness: is the collection underpinned by a sound selection policy to ensure comprehensive coverage. (IIPC Selection for Web Archives)?
8. Preservation of the blogosphere: does the repository succeed in capturing and rendering something of the whole extent, nature and context of the blogosphere?
9. Sharing and Interaction: can users instantly disseminate archived content using major social web platforms; and can they easily recommend new blogs for inclusion/archiving?
10. Meeting immediate user needs: do the archived blogs participate in the overall "scholarly record" [9], and how best to preserve this?

Out of scope are considerations of the different methods of harvesting / content creation between the three methods, which will not be explored in this study.

To ensure consistency of comparison across the platforms, a defined set of interesting and exemplary blogs has been selected, each of which is available for comparison in at least two of the platforms being studied.

4. PRELIMINARY CONCLUSIONS

Preliminary conclusions of the comparative study, are:

- That preserving parsed blog content (BlogForever) offers greater benefits in terms of discovery and fine-grained retrieval than preserving entire crawled websites (as per UK Web Archive)
- That websites stored in the WARC format (UK Web Archive) are more robust and better supported as coherent, preservable digital entities
- That PDF renditions of blogs (LSE) are easier and quicker to produce than using traditional web-archiving methods, but may in turn introduce additional preservation challenges
- That renditions of blog content viewed through the Wayback Machine (UK Web Archive) are perceived as more complete with regards to look and feel, attachments and layout than pre-processed renditions stored in XML (BlogForever)
- That a user-centric platform with tags, shopping baskets and other social media features (BlogForever) addresses the needs of user communities and curators more effectively than an inflexible and non-customisable view of the data
- That research value to scholars is enhanced by maintaining and indexing an aggregated collection of micro-detail from the blogosphere (authors, tags, comments)
- Aggregated collection of textual blog content will potentially be extremely useful to text-mining projects that are concerned with finding particular types of patterns, e.g. the evolution of language used on the internet, that cannot be easily discerned through the more usual title-based approach
- That XML-based blog content, capable of being exported into numerous library and metadata formats such as MARC XML, Dublin Core and METS, offers more flexibility for interoperability and sharing than WARC
- The three methods vary considerably in their searching facilities (speed of search, intuitiveness, interpretability of results)

5. ACKNOWLEDGEMENTS

Our thanks to the BlogForever project, the London School of Economics and the British Library for their assistance in conducting this survey using their materials.

6. REFERENCES

[1] Pennock, M. and Davis, R. 2009. ArchivePress: A Really Simple Solution to Archiving Blog Content. In: *Sixth International Conference on Preservation of Digital Objects* (iPRES 2009), 5-6 October 2009, California Digital Library, San Francisco, USA.

[2] Kim, Y., and Ross, S. 2011. Preserving Change: Observations on Weblog Preservation. In Proceedings of the 8th International Conference on the Preservation of Digital Objects (iPRES 2011)

[3] Stepanyan, K., Gkotsis, G., Kalb, H., Kim, Y., Cristea, A. I., Joy, M., Trier, M., Ross, S. 2012. Blogs as Objects of Preservation : Advancing the Discussion on Significant Properties. In: iPres 2012: Proceedings of the 9th International Conference on Preservation of Digital Objects.

[4] Consultative Committee for Space Data Systems, June 2012. Reference Model for an Open Archival Information System (OAIS), Recommended Practice CCSDS 650.0-M-2.

[5] Library of Congress PREMIS Editorial Committee, July 2012. PREMIS Data Dictionary for Preservation Metadata version 2.2.

[6] Library of Congress Network Development and MARC Standards Office, ND. Metadata Encoding & Transmission Standard (METS).

[7] Center for Research Libraries (CRL) and Online Computer Library Center, Inc. (OCLC), 2007. Trustworthy Repositories Audit & Certification (TRAC): Criteria and Checklist.

[8] Digital Curation Centre (DCC) and DigitalPreservationEurope (DPE), 2009. Digital Repository Audit Method Based On Risk Assessment (DRAMBORA).

[9] Hank, C. 2011. Scholars and their blogs: Characteristics, preferences and perceptions impacting digital preservation (Doctoral dissertation).Available from ProQuest Dissertations & Theses database (UMI No. 3456270)

Digital preservation of epidemic resources: coupling metadata and ontologies

João D. Ferreira
LASIGE, University of Lisbon
Portugal
joao.ferreira@lasige.di.fc.ul.pt

Catia Pesquita
LASIGE, University of Lisbon
Portugal
cpesquita@di.fc.ul.pt

Francisco M. Couto
LASIGE, University of Lisbon
Portugal
fcouto@di.fc.ul.pt

Mário J. Silva
INESC-ID, University of Lisbon
Portugal
mjs@inesc-id.pt

ABSTRACT

The preservation of epidemiological information is challenging in several aspects, since this is both a data-intensive and multidisciplinary subject, with large amounts of data spanning several domains of knowledge. We present, as a case study, the Epidemic Marketplace (EM), a platform dedicated to the preservation of epidemiological resources. To ensure integrity of the data, the EM uses a metadata model coupled with the Network of Epidemiology-Related Ontologies (NERO), a compilation of ontologies covering several domains of epidemiology. This enables users to quickly annotate their resources with concepts from those ontologies, increasing their visibility. Additionally, the ontologies of NERO offer support for future development, guaranteeing longevity of the metadata. This ensures that the information about the resources, such as its authors, is preserved and can be searched even in the absence of the data itself.

Categories and Subject Descriptors

H.3.7 [**Information Storage and Retrieval**]: Digital Libraries; H.3.5 [**Online Information Services**]: Data Sharing

Keywords

Data-intensive research, Digital Curation, Ontologies, Data sharing, Epidemiology

1. INTRODUCTION

Epidemiology research is a truly multidisciplinary subject as it relies on areas of knowledge as diverse biology, medicine, statistics, social sciences and geography. It requires, for instance, computational methods to predict the spread of a disease, realistic large scale models and automatic data collection. Only a framework able to accommodate these methodologies can ultimately deal with epidemiology.

Epidemiology is also highly data-intensive, making it a direct case supporting the fourth paradigm of science [9], which addresses the challenges raised from the need to validate, analyze, visualize, store, and curate the large amounts of generated data. For example, models of the spread of an epidemic disease rely on large amounts of information. This information can be very general and easily located (*e.g.* the size of the population being infected), or specific to the population and to the disease (*e.g.* the rate of contact between people). Finding this information in literature can be difficult, if not impossible, in the time frame of utility of those models. In fact, while some diseases have periodic surges, like the flu, with an expected number of peaks per year, other surges are more unpredictable, and modeling them in real time, as the disease progresses, requires the quick collection of the necessary information. The *E. coli* outbreak in Europe in 2011 is an example of such a situation.

Given these characteristics, it is particularly crucial to guarantee the preservation of epidemiological data, to ensure that it remains available, reliable and usable for the future [3]. If past data can be easily retrieved and explored, then the probability of it being reused increases, which is especially relevant in a complex domain such as epidemiology, where combining data across different locations, diseases or even time can lead to new insights and new knowledge.

2. THE EPIDEMIC MARKETPLACE

To answer to this need in information preservation, we have created the Epidemic Marketplace (EM), a platform for epidemic research that enables and encourages epidemiological data sharing, enabling the community to perform data-intensive research [11]. The EM was developed as part of the European Epiwork project, which aims at creating the appropriate framework of tools and knowledge needed for the design of epidemic forecast infrastructures to be used by epidemiologists and public health scientists [13]. It emphasizes the urge to share the information within the epidemiology community, and directly demonstrates the advantages of doing so. In fact the collaboration with other partners of the Epiwork project has shown the need to standardize the sharing of the digital resources.

3. THE EM METADATA MODEL

Metadata is an essential component of digital sharing and preservation, since it ensures that the data can be uniquely identified and accurately described to support future retrieval and reuse. As such, the EM establishes a metadata model to annotate its resources, based on the Dublin Core (DC). DC was chosen to be the base of this metadata model because it is an interoperable metadata standard, it provides a semantic vocabulary with many of the elements needed to manage an online datastore, and it enables a straightforward extension.

The metadata model defines the Network of Epidemiology-Related Ontologies, which contains concepts that are relevant for characterization of epidemiological resources. Our approach has the benefit of increasing interoperability with external services and, by restricting annotation to ontology-based controlled vocabularies, we move closer to the idea of a Web of Knowledge instead of a Web of Text [1].

The EM metadata model provides elements for (i) *technical information* (the uploader, an internal identifier and the date of submission); (ii) *general information* related to the digital resource, (*e.g.* title, creator – which need not be the same as the uploader); and (iii) *content-specific information*, such as the subject, the sources used by the resource or even the epidemiological information that makes up the resource.

The terms offered by the DC can already handle many of the requirements of the EM, especially in the *technical* and *general* information areas. However, epidemiology relies on multiple domains of knowledge. Accordingly, the metadata model devised for this purpose must extend the core elements of DC with tags appropriate for these domains of epidemiology. For example, many epidemiological resources deal with one or more diseases, a concept absent from DC; as such, the EM metadata model contains a specific element, `<em:disease>`, suitable for annotating a resource with the diseases it refers to. This property roughly translates to "the resource refers to disease X". Using metadata in this fashion ensures that the resource is searchable not only based on the general information provided by the DC but also based on its epidemiology-specific contents.

Furthermore, we extended some of the DC elements with new epidemiological elements. For example, the content-specific element `<em:biologicalInformation>` is refined by a number of biologically relevant elements, such as the previously mentioned `<em:disease>`.

Additionally, the metadata model specifies the expected values that can be used to fill each element. Some expect literal values, such as `<em:title>`, which expect a string. Most of the *content-specific* information must be selected from ontologies of an appropriate domain. For example, to fill the `<em:disease>` element, instead of the literal "flu", the URI `http://purl.obolibrary.org/obo/DOID_8469` should be used. This is the identifier of the concept named "Influenza" in the Human Disease Ontology. Several ontologies have been collected in a network of relevant ontologies, which have been integrated in the EM so that users of the platform can search them and correctly annotate their resources.

4. NERO

Most of the *content-specific* elements of the EM metadata model are filled with concepts from ontologies. To properly encourage users to annotate their data and ensure preservation, we integrated into the EM a number of ontologies that provide appropriate concepts that assist users during the annotation process. These were collected into a Network of Epidemiology-Related Ontologies (NERO) [8].

NERO directly contributes to the preservation of epidemiological resources in at least three ways:

1. its ontologies were selected in order to ensure both availability and longevity;
2. the meaning of the concepts is guaranteed to remain unchanged; if some modification happens, the concepts are marked as deprecated and a pointer to the new concept is made. This means that there will always be opportunity to update a deprecated annotation with the new term or to reconsider it;
3. as with any ontology, NERO allows the full spectrum of semantic web technologies to be used to search resources: it enables performing simple but powerful queries on the EM, or to draw inferences based on the semantics of these annotations [2], ensuring that pertinent data can be more easily retrieved and subsequently used, and thus fulfilling one of the main goals of digital preservation [3].

The ontologies contained in NERO were selected based on a number of requirements, some of which are related to the preservation of epidemiological resources. For example, these ontologies are required to provide textual definitions for their concepts, to be popular among the communities that use them and to be publicly available. All these properties contribute to the preservation of metadata integrity.

In our search, we found ontologies that already try to model the epidemiological domain. Given their low coverage and granularity, they were deemed inappropriate for inclusion in NERO. However, they provided a sense of the concepts that should be modeled in an epidemiological resource. Some general-purpose ontologies contain concepts of epidemiological interest. From a preservation point of view, these ontologies are adequate for annotation. However, properly scanning through these large terminologies and determining which of their concepts are relevant would be too colossal a task for the typical epidemiologist.

In face of these issues, we ended up selecting mainly single-domain ontologies for NERO. The OBO Foundry [14] defines a set of principles that must be fulfilled by an ontology before it can be included, enforcing good quality by promoting good practices in ontology development. In particular, its ontologies are public domain and must guarantee versioning, documentation, etc., which contribute to manageable preservation of their contents. Given their association to a high profile initiative, these ontologies are more likely to be kept available and up-to-date in the future.

5. INTEGRATION OF NERO IN THE EM

Annotation of resources with metadata is only effective if the users are encouraged to create this annotations. For this reason, there are two mechanisms in NERO that facilitate this process.

When users upload a resource to the EM, they are required to fill-in a minimal set of mandatory metadata elements, which include `<em:title>` and `<em:description>`. For the annotation of epidemiological information, the EM provides an autocomplete function that, based on user input, retrieves concepts from NERO which are appropriate for the metadata field in question. This effectively hides the technical details of the ontologies from the regular users, letting them focus on semantic annotation.

Additionally, each item in the list of suggestions is associated with its description, which users can read to help them choose the concept that better describes the resource. Whenever a given characteristic of the resource cannot be accurately described by any of the available ontology concepts, the user can easily assign a more general concept, which is supported by the inherent hierarchical nature of ontologies.

The second mechanism is the exploration of the actual resource provided by the user with text-mining to suggest back annotations that the user might think are relevant. This functionality uses NCBO BioPortal's Annotator service [10] to read the content of the resource, where possible, and preloads the annotation form with the NERO concepts it finds.

Given the variable nature of epidemiological resources, not all resources will need to be annotated in all metadata fields. For instance, a resource focused on tracking the geographical spread of a disease probably won't refer to any drugs, or if it focuses on the treatment of a disease, it might not include information on diagnostic method. In a recent analysis we conducted of semantically annotating over 100 Epidemiological resources in the EM, and found that all resources mentioned at least one disease and one geographical location, about 80% included information about the diagnostic method and the pathogen involved, but only about 30% mentioned any drugs or vaccines.

One crucial feature of the EM and NERO integration is the ability to assign multiple concepts to the same metadata field, since many resources mention multiple diseases, symptoms, drugs, etc., mirroring the wide scope of epidemiology. This effectively enables crossing information from different resources referring the same or similar entities, such as diseases or drugs, to support broader studies.

The adoption of a metadata model to support the semantic annotation of epidemiological resources, ensures a more structured annotation process, effectively guiding the annotation itself. Furthermore, by coupling the metadata model with NERO, the annotation process is further simplified, since terms to fill-in metadata fields are retrieved from a controlled vocabulary which is backed by the rich properties of ontologies such as hierarchical structure, definitions and other properties and relations.

6. BENEFITS FOR EPIDEMIOLOGY

Once epidemiological resources are annotated with NERO, metadata can be used in complex semantic analysis as part of diverse tasks, such as information retrieval and information extraction. These tasks will provide epidemiologists, particularly the modelers, with tools that enable an easy discovery of models and the data needed to parametrize them.

There are two main challenges in accomplishing this goal. The first is to define a way to effectively compare resources that are annotated using different sets of ontologies, *i.e.* how to compare a resource annotated with disease and pathogen, to another annotated with symptom and treating drug. This problem is relevant in the context of NERO, since different resources have different domains, and are, as such, annotated using different ontologies; and also because resources annotated with NERO may, at some point, be compared with resources annotated with other ontologies.

Semantic similarity [4,7,12] can address this issue. This will improve information retrieval by allowing a user to find resources that are similar to an input query. For instance, a user can be interested in finding all resources related to viral diseases. The system can retrieve resources similar to this query. Alternatively, the user can use as input a resource and find related ones. Semantic similarity across multiple ontologies exploits correspondences between the concepts, but such correspondences can be unavailable; ontology matching techniques can then be used to automatically create them, increasing the accuracy of similarity and, as such, the performance of information retrieval, and the field of ontology matching [5,6].

A second challenge resides in providing a contingency plan for handling cases where few or no annotations exist, which translates to how to generate annotations in an automated fashion for a given resource. Although we expect this situation to become increasingly less frequent as EM gains momentum, it will always remain a necessity to complement manual annotation.

Text mining is already used to handle this by extracting relevant information from the contents of EM resources, and then creating new annotations. This is particularly relevant in poorly annotated resources, since usefulness to the community directly depends on the ability to easily retrieve them. One of the main goals of such techniques is to facilitate the process of annotation to users. By analyzing the content of the files being uploaded, this techniques mine the data to find, for example, diseases or geographical places. These automatic annotations are suggested to the user, who can accept or reject them, improving the quantity and quality of annotations and contributing to a better performance in information retrieval.

When NERO ontologies do not have a sufficient degree of specificity, new concepts can be added using semi-automatic ontology extension, which is capable of automatically suggesting new concepts and relations. New concept suggestions can be derived from text or external ontologies and resources, or more interestingly from the free text annotations made by EM users.

The integration of these techniques into the EM will undoubtedly result in a full-fledged system for semantic web based information retrieval and extraction over its resources.

A final advantage of the EM is in the area of privacy. Since epidemiological data is generally sensitive, the EM manages data access in a fashion where, although the metadata is accessible by all users, access to the data itself can be protected and restricted to authorized users, ensuring that data can remain private, while some of the knowledge about the dataset is still shared, cataloged and found using automatic systems, contributing to its preservation and reuse.

7. CONCLUSIONS

In this paper we present the EM's metadata model, an extension to the Dublin Core coupled with a Network of Epidemiology-Related Ontologies (NERO). It was created with the aim of preserving digital epidemic resources. NERO was compiled based on a set of requirements that ensure, among other qualities, good preservation of the metadata. The EM metadata model supports the annotation of epidemic resources and the application of semantic web technologies over them. The integration of NERO with the EM made the annotation process easier and more complete, giving users a standard set of concepts to choose from. This has already resulted in a corpus of annotated epidemiological resources, over which the information retrieval and extraction system can operate.

By providing better tools to annotate the EM resources, these will be more easily preserved, guaranteeing an easier sharing of epidemic resources for the foreseeable future. In fact, NERO is able to serve all the epidemiology community, since it is not bound to the EM but can subsist on its own. For example, research teams working on developing approaches to identify and quantify modularity in spatially structured and heterogeneous meta-populations and contact networks can also benefit from using NERO, both as an annotation standard and as a way to search for other resources. The geospatial information that NERO encodes can be of great interest here. The collection of validated data through ICT applications can also benefit from NERO: semantically annotating these data is a major step in its analysis, and NERO can serve as the source of concepts for this annotation. The establishment of this network of ontologies contributes, therefore, to an improvement for all the community, particularly on the topics of preservation, sharing and reusing epidemiological data.

8. ACKNOWLEDGMENTS

The authors want to thank the European Commission for the financial support of the EPIWORK project under the Seventh Framework Programme (Grant #231807), and the Portuguese FCT through the financial support of the SOMER project (PTDC/EIA-EIA/119119/2010) and the PhD grant SFRH/BD/69345/2010, and the PIDDAC Program funds for INESCID (Pest-OE/EEI/LA0021/2013) and LASIGE multi annual support.

9. REFERENCES

[1] T. Berners-Lee and J. Hendler. Publishing on the semantic web. *Nature*, 410(6832):1023–1024, 2001.

[2] C. Bizer, T. Heath, and T. Berners-Lee. Linked data-the story so far. *International Journal on Semantic Web and Information Systems (IJSWIS)*, 5(3):1–22, 2009.

[3] E. Conway, D. Giaretta, S. Lambert, and B. Matthews. Curating scientific research data for the long term: a preservation analysis method in context. *Work*, 6(2):38–52, 2011.

[4] F. M. Couto, M. J. Silva, et al. Disjunctive shared information between ontology concepts: application to Gene Ontology. *J Biomed Semantics*, 2(5), 2011.

[5] D. Faria, C. Pesquita, E. Santos, F. M. Couto, C. Stroe, and I. F. Cruz. Testing the AgreementMaker System in the Anatomy Task of OAEI 2012. *arXiv preprint arXiv:1212.1625*, 2012.

[6] J. D. Ferreira, D. S. Batista, F. M. Couto, and M. J. Silva. The Geo-Net-PT/Yahoo! GeoPlanet (TM) concordance. *Technical Report. TR 10-05, Departamento de Informática da Faculdade de Ciências da Universidade de Lisboa (doi:10455/6677)*, 2010.

[7] J. D. Ferreira and F. M. Couto. Semantic similarity for automatic classification of chemical compounds. *PLoS computational biology*, 6(9):e1000937, 2010.

[8] J. D. Ferreira, C. Pesquita, F. M. Couto, and M. J. Silva. Bringing epidemiology into the Semantic Web. In *Proceedings of the International Conference on Biomedical Ontologies*, 2012.

[9] J. Gray. Jim Gray on eScience: a transformed scientific method. *The fourth paradigm: Data-intensive scientific discovery*, 2009.

[10] C. Jonquet, N. Shah, C. Youn, C. Callendar, M. Storey, and M. Musen. Ncbo annotator: semantic annotation of biomedical data. In *IntâĂŹl Sem Web Conf (ISWC)*, 2009.

[11] L. Lyon, A. Ball, M. Duke, and M. Day. Developing a Community Capability Model Framework for data-intensive research. In *iPres 2012-9th International Conference on Preservation of Digital Objects*, pages 9–16, 2012.

[12] C. Pesquita, D. Faria, H. Bastos, A. Ferreira, A. Falcao, and F. Couto. Metrics for GO based protein semantic similarity: a systematic evaluation. *BMC bioinformatics*, 9(Suppl 5):S4, 2008.

[13] M. J. Silva, F. Silva, L. F. Lopes, and F. M. Couto. Building a digital library for epidemic modelling. In *Proceedings of ICDL*, pages 23–27, 2010.

[14] B. Smith, M. Ashburner, C. Rosse, J. Bard, W. Bug, W. Ceusters, L. J. Goldberg, K. Eilbeck, A. Ireland, C. J. Mungall, et al. The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. *Nature biotechnology*, 25(11):1251–1255, 2007.

Risk Management for Digital Long-Term Preservation Services

Stefan Hein
German National Library
Adickesallee 1
D-60322 Frankfurt am Main
+49-69-1525-1722
s.hein@dnb.de

Karlheinz Schmitt
German National Library
Adickesallee 1
D-60322 Frankfurt am Main
+49-69-1525-1782
k.schmitt@dnb.de

ABSTRACT
This article presents an ingest level system which has been developed as part of the Digital Preservation for Libraries (DP4lib) project. The purpose of the system and its implementation is to facilitate automatic technical quality checking of digital materials. It represents an essential part of the risk management system within the long-term preservation processes of the German National Library (DNB). Initial practical experience is reported upon, demonstrating that a significant step has been taken towards ensuring the long-term usability of digital materials.

Categories and Subject Descriptors
Standardization, Verification

General Terms
Management, Reliability, Verification.

Keywords
Digital Preservation, Risk Management, Ingest-Level, Quality Management

1. INTRODUCTION
Handling risks is part of the daily business of long-term digital preservation. In all the areas of long-term digital preservation examined here, it is always important to recognise risks at an early stage, to assess their possible effects, to develop countermeasures and to implement these as required. Such risk management in organisations must be institutionalised in order to ensure continual monitoring of potential risk sources and to minimise any impact.

But how can comprehensive risk management be achieved for long-term digital preservation and its operational processes?

Risk management in this context is often referred to in the literature, e.g. in the OAIS reference model [1], as an integral part of preservation planning. The primary purpose of risk analysis in the ExLibris Rosetta system is to warn against the threat of obsolete file formats [2]. There it is carried out by the repository manager and is based on the data currently being managed. The approach presented here, by contrast, is distinguished by proactive measures taken right from the point at which the digital publication is ingested and regards any "inferior object quality" apparent at this time, which is based on more than an analysis of the file format, as a risk for future preservation action.

The objectives of the two-year Digital Preservation for Libraries[1] (DP4lib) project launched by the DFG were to evaluate the possibility of setting up a long-term preservation service for third parties and to implement a prototypical solution. An overview of the project results can be found in the long-term digital preservation manual [3] for service providers and users. Reflecting the main results of the project, one of the main benefit was that a suitable system of a cooperative risk management was set up consisting of automatic technical quality checking of digital objects and full reporting of all long-term digital preservation activities. The purpose was to lay the foundations for a trusted repository.

One of the main sources of risks in long-term preservation lies in the digital materials to be archived. The technical quality of the digital materials, for instance, is often both unknown and substandard, meaning that preservation of their long-term usability is already questionable with our current knowledge.

To check and if necessary avoid such risks, the service users and providers must cooperate to set up a joint risk management system which can recognise risks at an early stage and avoid them if possible.

The key component of the risk management ingest level system is described in section 2. Section 3 focuses on the technical implementation. The ingest level system ensures that risks associated with the partnership on the one hand and on the wide range of file formats on the other can be automatically recognised and communicated. The initial practical experience is presented in section 4. Finally, the last section includes a summary and the outlook for the further development of this approach.

2. THE INGEST LEVEL SYSTEM
The idea behind the ingest level system is presented in this section. The ingest levels are first defined and then the organisational integration and the contribution to risk management within the DNB are examined. The DNB actively uses the ingest level system for its internal long-term preservation processes, for ingesting digital publications as well as for the planned long-term preservation service for third parties.

The idea of using different levels for controlling and checking within long-term preservation is not new. Within PREMIS, for instance, different preservation level types were introduced which are based closely on groups of significant document properties which need to be preserved [4]. As in the ingest level system, preservation of the bitstream constitutes the first level. A similarly

[1] Project homepage: http://dp4lib.langzeitarchivierung.de/

close connection between level and preservation strategies can be found in the DHEP project [5] in which a total of 4 different levels of preservation strategies were introduced. By contrast, the ingest level system concentrates exclusively on checking the technical quality of a range of file formats and provides an indication of possible risks for the long-term usability of digital documents.

2.1 Definition and criteria

Assignment to an ingest level is the result of a tiered automatic checking process for file formats which is carried out (in part) in cooperation between the DNB and the depositing partners. By assigning an ingest level to a digital publication qualitative statements can be made about certain technical aspects of a digital object. A technical quality standard can also be expressed for the publication.

The general goals of this quality check, which is to be run for each file in each ingest transaction, are safeguarding the authenticity of the digital objects received and carrying out an analysis aimed at recognising technical restrictions at an early stage which hinder or even prevent the task of long-term preservation and also use of the digital objects.

Five test criteria, each one following on from the next, have been defined for this purpose:

1.) File integrity (DI)

The files submitted by the depositors have not changed during the course of the data transfer and processing.

2.) Identification (ID)

The file formats of the digital publication's files have been clearly identified.

3.) Lack of restrictions (LR)

The file object is free of restrictions, i.e. there are no recognisable (to the DNB) technical barriers which could impede or prevent the use or long-term preservation of the publication.

4.) Extraction of format-specific technical metadata (MD)

Format-specific metadata which are required for digital preservation could be generated.

5.) Format validity (V)

The file format (specifications) of the publication is valid.

Table 1 shows how the individual criteria relate to each other.

Table 1: Ingest level and criteria

	DI	ID	LR	MD	V
Level 0	X	O	O	O	O
Level 1	X	X	O	O	O
Level 2	X	X	X	O	O
Level 3	X	X	X	X	O
Level 4	X	X	X	X	X

Following the technical test, a digital publication is assigned level 0 if the integrity (DI) of the files belonging to the publication could be checked, confirmed and logged following the successful transfer to the DNB as the result of coordinated processes between the depositing institution and the DNB. Special procedures (checksum tests) are used for this. A digital publication is then assigned ingest level 1 if the file format could be successfully identified. No restrictive mechanisms may be detected in the subsequent analysis of the digital publication which impede or prevent the use or functionality of the publication for the issue of the next ingest level (ingest level 2). In the case of PDF documents, these include e.g. password, copy or printing restrictions which would prevent the issue of this ingest level. Ingest level 3 is assigned if sufficient additional format-specific technical metadata for long-term preservation measures could be extracted. The DNB has specified a core set of technical metadata for each file format. Currently the highest, and therefore the "best", level (ingest level 4) is achieved by digital publications if the validity of the file format used could also be positively tested.

The higher the ingest level, the more criteria have been positively tested and therefore the greater the risk management probability that the deposited publication can be preserved.

This form of technical qualitative analysis allows the DNB, for the first time, to automatically recognise long-term preservation risks for digital publications and to undertake suitable countermeasures at the time of transfer. As a consequence, the question arose as to whether countermeasures should be taken as a suitable response to the identified risks - and if so, which. The DNB has drawn up a format policy for the ingest and processing of digital publications.

2.2 Format Policy

A list of the minimum and maximum ingest levels for the file formats has been drawn up for the file formats deposited at present with the DNB on the basis of the current technical analysis possibilities. Table 2 contains an extract from this list. By setting a minimum quality standard for archivable file objects it was possible to draw up a format policy which contains rules for accepting and rejecting digital publications and also provides rules for further analysis tasks.

Table 2: DNB Format-Policy.

File Format	Min. ingest level	Max. ingest level
PDF	2	4
EPUB	2	4
…	…	…

The ingest of a publication is rejected on technical grounds if an ingest level below 2 is determined for a file of the digital publication. In such cases the DNB contacts the depositor. All other publications assigned an ingest level of 2 or higher are accepted into the archive system of the DNB. If some of the publication files have only been assigned ingest level 2 or 3, this does not constitute grounds for rejecting the publication. With regard to long-term digital preservation, the DNB is responsible for preserving the individual files of the publication in a permanently usable state and for carrying out any necessary preparatory measures.

The ingest levels are henceforth to be interpreted as new minimum expectations for the assessed quality standard of the individual file formats in the import process.

3. TECHNICAL IMPLEMENTATION AT THE DNB

The following section describes the technical implementation of the approach for risk management based on the DNB's ingest process for digital publications.

As shown in Figure 1, the DNB ingest process starts with the deposit of the digital publications via mass deposit interfaces such as OAI-PMH. It also includes the import processing chain for storage in the repository and ends with a further workflow, independent of the import process (LtpBinding), for transfer to the Long-term archive (LTA). The main steps of the risk management-enhanced import process include tasks such as checking for duplicates, issuing persistent identifiers, carrying out checksum checks, generating technical metadata and conducting the ingest level comparison.

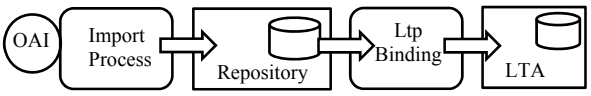


Figure 1 The Ingest Workflow.

3.1 Checksum test

The checksum test is one of the first test routines in the DNB import process; the first step involves calculating a checksum at the file level. This is then compared with that calculated and supplied by the depositor. Only if both checksums concur will the file object be assigned ingest level 0 and be forwarded for further processing. Ingest level 0 therefore constitutes the basis for all other process stages shown in Figure 2. At the DNB these are contained in a tool called *diagnose digital objects (didigo)*.

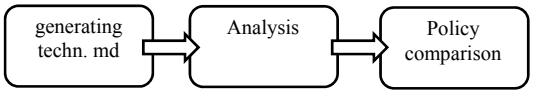


Figure 2. Diagnosis of digital objects.

3.2 Generation of technical metadata

For some time now the automatic generation of technical metadata using metadata tools has been a recognised and established component of the ingest process. The DNB has long been using the *File Information Tool Set (FITS)* as a framework for using an entire tool set. This framework provides access to a whole range of tools including the *JSTOR/Harvard Object Validation Environment (JHOVE)* tool, the *Digital Record Object Identification* (*DROID*) tool and the *NLNZ Metadata Extractor*. JHOVE cannot handle the same variety of file formats as DROID, however it does support the generation of technical metadata and also checks the formal accuracy and format validity. DROID, by comparison, merely identifies the file format and its version. Use of a tool set widens file format support and reduces the risk of errors in the identification and validation of the file format. FITS also offers significant added value in the form of easily configurable standardisation of the different tool outputs into the FITS format using XSLT. The DNB has used this function to adapt the FITS output to its own requirements, e.g. incorporating other metadata elements not included in the FITS distribution into the standardisation. However, the resulting output schema still complies with the FITS standard. This extended FITS format provides a format-specific metadata set which unifies the different technical metadata elements of a number of metadata tools and combines them structurally into a single standard [7]. A further adjustment which the DNB has made is the integration of a DNB tool to analyse files in ePub format.

3.3 Analysis

The FITS processing is followed immediately by analysis of the results. This is concluded by final calculation of the ingest level which is initially set at 0. The test criteria of restriction-free access, file format, format-specific metadata and format validity are examined - in this order - on the basis of the FITS output. Each test which is successfully passed raises the ingest level incrementally by 1, with 4 being the highest ingest level achievable by a file object. As soon as one of the above tests has been failed, the ingest level remains at its present level.

FITS yields XML objects, meaning that the technical implementation of this test can consist in querying individual XML elements using e.g. XPATH expressions. An example here is the corresponding expression for the file format test criterion:

```
/fits:identification[@status='UNKNOWN']
```

This expression checks the existence of the kind element *identification* which has the attribute *status* and the value *unknown*. The existence of such an element indicates that FITS was not able to identify the file format. This means that the test criterion for granting ingest level 1 has not been met. As noted above, the incremental increase in the ingest level stops here and the ongoing results analysis is discontinued. The file object is forwarded marked ingest level 0 to the next stage, the ingest level comparison.

3.4 Ingest level comparison

The depositor-dependent format policy is loaded for the ingest level comparison. This sets the minimum ingest level to be reached for each file format. The relation between file format and ingest level is established using the *PRONOM Unique Identifier (PUID)* issued by DROID. For example, if the definition of ingest level 2 is reached for PUID fmt/16, only file objects in PDF format version 1.2 for which

- the bitstream passes the integrity test
- the file format is identified and
- no use restrictions apply

will be ingested into the DNB repository and therefore into the preservation repository. If a publication consists of multiple files, all its elements must meet the set criteria, with the lowest value determining the overall ingest level.

4. PRACTICAL IMPLEMENTATION AND EXPERIENCE

Following on from the description of the basic idea and technical implementation of the risk management issues, the intention below is to present an overview of the experience gained to date.

The system was put into operation in December 2012 as part of the DNB operational processes for handling digital publications. The vast majority of files undergoing the risk management processes since then have been PDF and ePub objects. Figure 3 (date: 12.4.13) shows the distribution of analysed ingest levels. The visualised results show the figures for file objects submitted to the DNB which fulfil the requirements of the DNB internal format policy.

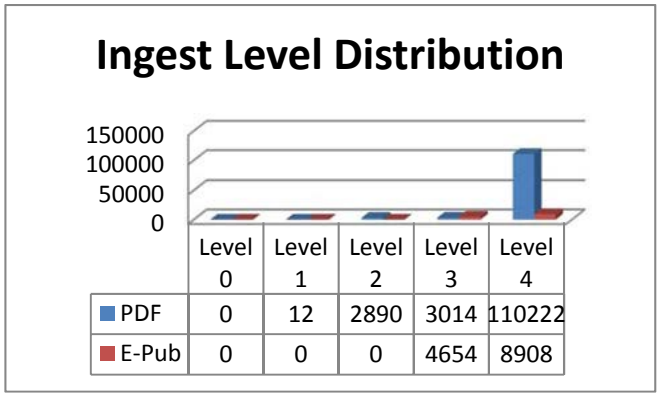


Figure 3 Ingest level distribution (PDF, ePub) in the period from 12/2012 to 04/2013

Of the total of 116,138 PDF files, the vast majority (110,222) are ingest level 4. Even though only 3,014 PDF objects had a validity problem (ingest level 3) and no technical metadata could be generated for 2890 objects (object level 2), this absolute figure is likely to rise and should not be underestimated. Several thousand problematic objects can accrue within just a few years; these need to be prioritised for preservation strategies such as format migration. With regard to the ePub format, only half of all the objects transferred to date are free of validity problems.

It should also be mentioned that the fact that a clear majority of objects are ingest level 4 does not necessarily signify that this majority automatically represents the "object quality" of the publication world. It should be borne in mind that only through the risk management measures and the resulting requests by staff for "better" versions were many objects of ingest level 3 or lower able to be raised to ingest level 4. The return of defective objects raises awareness amongst the publishers of the need to attach greater importance to the quality of their objects. In some cases this has already led to checking tools being integrated in the publishers' production processes. Despite all the automation systems, these costs associated with manual risk management activities, including e.g. necessary adaptations to the format policy, should not be neglected in any cost assessments.

4.1 Technical limits

In many cases, file objects which only achieve ingest level 2 reveal their technical limits in the validation tools used. At present, for example, some PDF variants (e.g. PDF/X) cannot be correctly processed, meaning that the resulting technical metadata deficiencies are not always due to supposedly "poor" object quality.

A clear discrepancy between theory and practice has also emerged in format validity. The differing interpretations of the HTML standards by the panoply of disparate browser providers and the resulting differences in the ways in which a website are displayed are acknowledged examples of this. Additionally, the library's ePub-Analyzer metadata tool which checks conformity of ePub files against the ePub specifications often identifies a lack of schema validity in the toc.ncx file which describes the table of contents. However, practical tests of their display and use on current devices showed that this validity problem is negligible at present. Nevertheless, from the perspective of long-term preservation it represents a significant risk factor which can be dealt with in the preservation strategy planning e.g. by means of suitable corrective measures.

A total of 12 individual ingest level 1 PDF objects are shown in Figure 3, some of which are attributable to different results obtained by the tools operating in FITS with regard to the existence of usage restrictions. In these cases, manual analysis showed that use of the objects was not restricted.

Finally, the ongoing development of file formats for electronic publications poses further demands in terms of constant updating and development of the metadata tools used. During transition periods in which tool support is still incomplete, compromise solutions, e.g. lowering of the ingest level, should be considered.

5. Summary and outlook

The present article examines the DP4lib ingest level system and its practical use in the DNB. This system introduced automatic quality checking to the DNB's long-term preservation activities as part of a comprehensive risk management system. It was shown that risks which are ubiquitous in the file formats of digital materials can be detected and classified at an early stage. The first countermeasures designed to reduce file format risks were the formulation of a format policy and the setting of a limit beyond which the task of ensuring the long-term usability of digital objects can no longer be fulfilled. Initial experience shows that the automatic quality analysis has yielded accurate findings regarding the technical quality of the library's stocks. The data can also be used as the basis of improvement processes and to reduce long-term preservation risks. The ingest level system therefore provides a practicable control instrument based on tangible limits and rules of action. It also allows depositing partners to formulate their own requirements and expectations in terms of object quality and risk analysis, thereby facilitating the creation of service agreements between DP4lib service users and providers. It should be added that this approach has also resulted in a number of terms entering the vocabulary of the specialist and IT departments of the DNB, leading to a corresponding improvement in communication.

In the future it should be established whether the five levels (and their order) in the current ingest level system and the related weighting are sufficient to address the long-term preservation risks for digital publications and the associated problems arising from the growing variety of file formats.

6. References

[1] The Consultative Committee for Space Data Systems (CCSDS), June 2012. *Reference Mode for an Open Archival Information System (OAIS), Recommended Practice.*

[2] Ex Libris Group, 2010. *Ex Libris Rosetta: A Digital Preservation System – Product Description.*

[3] Langzeitarchivierung – Ein Handlungsleitfaden für Dienstleister und Dienstnehmer. http://dp4lib.langzeitarchivierung.de/

[4] PREMIS With a Fresh Coat of Paint http://www.dlib.org/dlib/may08/lavoie/05lavoie.html

[5] Data Preservation in High Energy Physics; David South; Proceedings of plenary talk given at the 18th International Conference on Computing in High Energy and Nuclear Physics (CHEP 2010).

[6] *File Information Tool Set;* http://code.google.com/p/fits/

UPBox and DataNotes: a collaborative data management environment for the long tail of research data

João Rocha da Silva
INESC TEC, FEUP,
Universidade do Porto
Rua Dr. Roberto Frias, s/n
4200-465 Porto Portugal
joaorosilva@gmail.com

José Pedro Barbosa
FEUP, Universidade do Porto
Rua Dr. Roberto Frias, s/n
4200-465 Porto Portugal
ei08036@fe.up.pt

Mariana Gouveia
FEUP, Universidade do Porto
Rua Dr. Roberto Frias, s/n
4200-465 Porto Portugal
ei10124@fe.up.pt

João Correia Lopes
INESC TEC, DEI, FEUP,
Universidade do Porto
Rua Dr. Roberto Frias, s/n
4200-465 Porto Portugal
jlopes@fe.up.pt

Cristina Ribeiro
INESC TEC, DEI, FEUP,
Universidade do Porto
Rua Dr. Roberto Frias, s/n
4200-465 Porto Portugal
mcr@fe.up.pt

ABSTRACT

Current research data management workflows are often an *a posteriori* process, with research datasets being targeted for preservation actions after the whole research process is completed. This approach works well for research publications but not for research datasets due to their dynamic nature. It is important to gather data production contexts, so the data management process should be present since the start of the research, effectively becoming a part of the workflow. Due to their rigid workflow-based deposit approach, widely used repository solutions are not intended to support the fast-paced evolution of datasets as they are produced. In this paper, we present a collaborative data management environment designed to help a small research group store and describe their datasets in preparation for later deposit in a data repository. It is built on two integrated, open-source components: UPBox—a private cloud and web-based file storage environment—and DataNotes—a solution tailored for researchers to collaboratively describe their data, based on Semantic MediaWiki. Preliminary usage tests have shown that the features of this solution respond to data management needs in research groups.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Digital Libraries; H.4 [**Information Systems Applications**]: Miscellaneous

General Terms

Human Factors, Management, Standardisation

Keywords

Research data management, data repositories, Semantic MediaWiki, digital curation

1. INTRODUCTION

Research data management is assuming an increasingly relevant role in the research workflow. The adoption of appropriate research data management practices presents advantages for research funding institutions (e.g. international recognition of their project's results) but it is ultimately the researchers who must realise the potential improvements to their work that may come from the adoption of such practices. These have already been extensively discussed and include increased citation rates for articles that provide access to base data, reproducibility of research results, formulation of new research questions [10, 3, 7] and also the wider goal: faster advancement of science [2]. These goals, while important, are often seen by researchers as unclear long-term benefits of a process that requires a substantial time investment on the researcher's part.

Current research data management workflows usually rely on a dataset description process performed by professional curators. While this process is effective for producing high-quality generic metadata, the inclusion of domain-specific metadata in the description of research datasets requires the close collaboration of the dataset creators, which are experts in the domain but often lack the data management skills required to perform comprehensive descriptions of their datasets [13]. Only through this cooperation can we produce rich domain-specific descriptions for research datasets [6]. However, this approach tends to require too much time from researchers, who often do not realise any immediate advantages in the data management process. At the same time, data curators become the bottleneck in the curation process—the end result is a process that can turn into a series of sporadic contacts and lost opportunities for describing those datasets as their authors move to pursue other research questions.

While community-supported research data repository direc-

tories are already a reality—an example is DataBib, a directory for research data repositories [14]—collaborative environments for curators and researchers to describe datasets are still in their early stages. In 2013, the DataUP project [12] has shown how a self-deposit tool built directly into Microsoft Excel can help researchers deposit spreadsheets directly from their working environment. An interesting aspect of the project is that it focuses on guiding researchers through the description of the spreadsheets, pointing out possible mistakes in their formatting and organisation, while making it easier to describe them using standardised metadata.

With this work, we present an approach at data management that has the primary goal of making it an ongoing process that supports the everyday activities of a researcher—a view that has already been expressed in a recent report [5]. This more dynamic environment relaxes some interoperability requirements and strict metadata production workflows in favour of capturing the data and its context as it is produced and processed. At the same time, it provides a set of features that immediately reward researchers for their efforts in describing and organising the datasets. By using UPBox—a "Dropbox" for research datasets—and DataNotes—a data description wiki—to manage the data, they gain access to a safe and simple file storage area for the datasets, easier data sharing within their research group and a collaborative wiki-based data description tool for the datasets.

Since researchers can be reluctant to deposit research datasets on infrastructures outside of their control, we have designed this environment to work completely under the research institution's control, that is, in its own servers. We see this environment as a "staging area" to prepare datasets for later repository ingestion. An important part of the data description work will already be done by the time the final data is produced and research results published, effectively making it an easier task and encouraging researchers to complete the data management process with the assistance of data curators. This work is oriented towards the "long-tail of data" as we are not trying to manage the very large datasets created in some research areas—which are often supported by appropriate infrastructures—but rather the myriad of small datasets produced by diverse research groups [9], which tend to be more at risk due to the scarce data management resources available in their projects.

2. COLLABORATION: THE KEY FOR USER ENGAGEMENT

Most current research data management workflows take an *a posteriori approach*. This means that researcher involvement in the process is reduced to a certain point in time when the datasets are curated and deposited in a repository platform. Some advantages of the *a posteriori* process are its simplicity in terms of planning (for both researcher and host institution), a relatively reduced effort and easy learning curve for researchers. More importantly, it yields comprehensive, standardised dataset metadata. However, our past experience has shown that this approach also has some drawbacks concerning the number of datasets that are actually preserved.

Another issue surrounding dataset description is *timing*. Interaction between researchers and curators usually takes place at a relatively late phase of the research activities, after researchers have gathered and processed their datasets, obtained results and published them. While currently this is the most common practice in publication management, it is clear that research data curation should start as early as possible in the research process [8, 5]; datasets should be described as soon as the researcher possesses adequate domain knowledge and has created them, since that is when the data production context is completely available [1].

In 2012, the UPData project [11] provided insight on the features that researchers find interesting in a data management workflow. Ensuring the reproducibility of research findings and relating publications to their base data is interesting, but researchers tend to focus on more immediate benefits of integrating the datasets in the research data management workflow. Among these are, for example, easy sharing among research colleagues—sending an URL to a resource where the dataset is available is a basic but clear example. One of the main reasons that make researchers reluctant to produce metadata for datasets is the work involved in filling in descriptors that they often see as irrelevant in their own domain. To make the process less tedious (albeit with a compromise in interoperability) metadata schemas can be replaced with application profiles: "schemas which consist of data elements drawn from one or more namespaces, combined together by implementors, and optimised for a particular local application" [4]. It is hard, however, for curators to manually design specific application profiles on a research group/project basis, so these profiles should emerge naturally through descriptor reuse on each domain—a collaborative description environment is a pre-requisite for this to happen.

In most cases, research activities are performed by groups of researchers in close collaboration, so it makes sense to reduce duplicate efforts and make research data management a collaborative effort as well. In fact, data management can even be useful to research teams by helping them share data within the research group, while encouraging the team to share description efforts as well. Metadata production in a collaborative context becomes *rewarding in the short term*, allowing the data management environment to become a central hub of the research activities. As a side effect, application profiles may surface as the descriptors from different metadata schemas are reused in different domains.

3. COMBINING A PRIVATE CLOUD WITH A SEMANTIC WIKI

Our proposed research data management environment is built on two main components, interconnected by a set of web-based communication endpoints or *web-services*.

The architecture of the system is shown in Figure 1. UPBox (1) uses the server's local storage, which can be mapped to a RAID-based storage (our selected solution), network volume, or distributed storage layer. A possible alternative would be a Hadoop File System (HDFS) mountable volume[1] to provide abstraction over a private cloud for hor-

[1] `http://wiki.apache.org/hadoop/MountableHDFS`

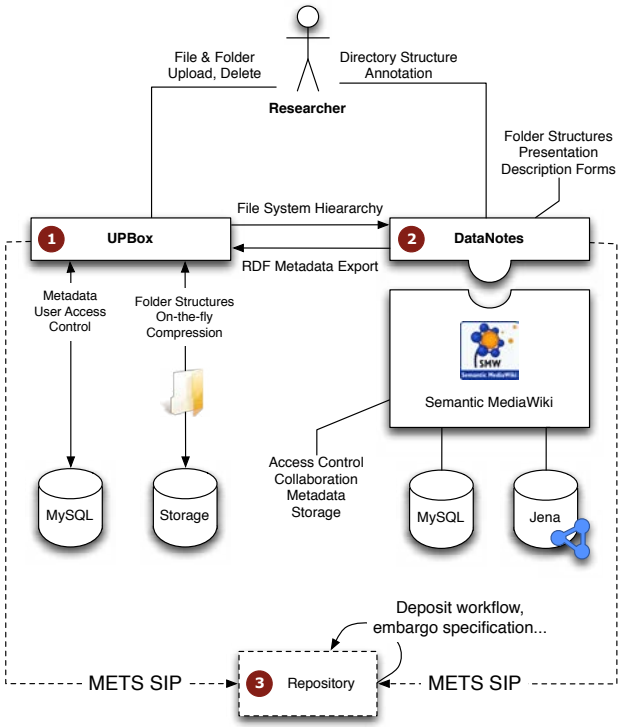


Figure 1: Architecture of UPBox and DataNotes

izontally scalable storage. A MySQL database is used to save the data required for user management, access control and metadata concerning the directory structures. All files are compressed and decompressed *on-the-fly* when users upload/download them to/from the server, to minimise the storage space required to support the system. UPBox is connected to U.Porto's central information system (SIGARRA) via an LDAP (Lightweight Directory Access Protocol) plugin, enabling U.Porto staff to log into the system using their SIGARRA credentials. External users can also register in the system, enabling inter-university collaborative work. The platform allows researchers to create "projects", areas where folders can be created and files can be deposited much like *Dropbox* folders. A project can be shared with team collaborators by adding members (the system provides suggestions from the list of registered users). Members of a project can upload files, as well as create folders or delete them. Several files can be uploaded simultaneously to facilitate the migration of existing datasets.

DataNotes (number 2 in Figure 1) is a wiki-based directory structure annotation platform, built on top of Semantic MediaWiki[2]. It allows users to quickly produce wiki pages containing the metadata for their datasets. The goals fulfilled by DataNotes are:

1. Providing a collaboration environment for describing directory structures, supporting version control, locking, concurrent edition management, namespaces and user access control.

2. Helping researchers in the group to find datasets via text-based search over the metadata.

3. Offering a friendly user interface, albeit with sophisticated capabilities to capture relationships between parts of the dataset and also semantic inter-dataset links for those cases where such detail is required.

4. Easy sharing of dataset descriptions (ideally the ability to send a direct link to a described folder or file).

5. Absence of dependencies on closed source solutions, modules or libraries that may endanger the access to the data stored in the solution as it becomes deprecated and there is no way to update or review its business logic.

6. Ease of installation, making it easy for any research institution to host their own DataNotes instance to support the work of their research groups.

7. Preparing datasets for long-term preservation by easing the export of dataset metadata records in a standard format (e.g RDF), ensuring the survival of the data even in the event of DataNotes being replaced with another platform.

8. Providing programmatic search capabilities that enable resource retrieval from the wiki, based on criteria specified by external systems.

Since DataNotes is based on a wiki platform, namespace management and access control features are already present, along with concurrent editing capabilities and continuous versioning of the wiki pages which contain file and folder metadata. Free text search is also present, allowing users to retrieve dataset pages via a global search function. The interface can be considered user-friendly as most of the standard MediaWiki components are maintained in Semantic MediaWiki and remain unchanged—keeping the easy learning curve that continues to make it possible for non computer-savvy users to write and review Wikipedia pages. The system also allows users to share dataset descriptions easily, since every description is a wiki page with its own unique URL—these URLs are shown in the web browser during navigation and can simply be copied and pasted in a message for sharing with other users that have permissions to access the resource.

The "Repository" module (number 3 in Figure 1) represents an existing repository (such as DSpace). After datasets are deposited in UPBox and their descriptions produced in DataNotes, researchers should be able to automatically package the existing state of a folder (for example) and send it in to the repository, where a new ingestion workflow will be started. The metadata for the new dataset will be subjected to all the usual validations by a curator (including embargo specifications) and then deposited in the repository. At that time, and due to the "static" nature of the resulting repository resource, it can be cited safely in publications via a persistent identifier (URI).

[2] `http://semantic-mediawiki.org`

4. CONCLUSIONS AND FUTURE WORK

Observation of current practices with research data suggests that data management should accompany researchers in their everyday activities instead of being performed *a posteriori*. The goal is to maximise the opportunities for gathering datasets, allowing their later ingestion into a repository for long-term preservation. Another goal is to make it possible for institutions to maintain complete control over the data produced by their researchers. To address these needs, we have designed and constructed a fully open-source collaborative environment for data sharing and description among research groups. The system allows researchers to deposit their datasets in Dropbox-like folders and then describe them using an integrated wiki interface.

Presently, there is no support for versioning in the UPBox platform—unlike DataNotes, which already offers versioning capabilities for the metadata pages of each file/folder since it is built on Semantic MediaWiki. Disk quotas for UPBox users are also in the list of possible improvements to provide control over the server's storage space. A more sophisticated access control system could also be implemented, allowing project owners to specify the actions to be performed by each team member for each folder (and subfolders) in the project. An UPBox desktop client to enable seamless synchronisation with the remote storage (much like *Dropbox*) is also planned, as well as a folder upload feature, to make it easier to migrate an entire existing directory structures to UPBox.

The possible improvements for DataNotes have to do with making dataset description easier by automating repetitive tasks. By allowing researchers to reuse sets of descriptors from a folder to annotate another, we can encourage the creation of application profiles for each domain through community reuse. Also, to complete the data management cycle, datasets described in this environment should be handed over to a repository in a transparent way, at a moment chosen by the project owner. To achieve this, a connection to a data repository must be available, and we plan to use DSpace to build upon previous work on DSpace extensions for managing research datasets. In the future it will be possible to start DSpace deposit workflows directly from UPBox or DataNotes; given DSpace's OAIS-compliant endpoints, these systems must be capable of building METS SIP packages and submitting them to DSpace. Our goal is to make this process fast enough for researchers to be able to cite their datasets at the time of the publication of results, making it easier for their audience to find the corresponding base data.

A small validation experiment with a group of researchers from the FEUP Mechanical Engineering department was performed; thus far, the feedback on the improvements introduced by this platform has been positive, and has provided some insight on further development. For example, the decision to allow external users to register in the system was taken due to the fact that this research group included members from UTAD (University of Trás-os-Montes and Alto Douro), and they wanted to use UPBox to share datasets in the group—a situation that we found very likely to occur in the future. As the tools start to be used by different research groups, we will also determine if these tools should act only as a "staging area" or if they should be extended to satisfy long-term preservation requirements as well.

5. ACKNOWLEDGEMENTS

This work is supported by research grant "SFRH/BD/77092/2011", provided by the FCT—Fundação para a Ciência e Tecnologia (Portuguese Foundation for Science and Technology) and by the ERDF—European Regional Development Fund through the COMPETE Programme (operational programme for competitiveness), supported by National Funds through the FCT—Fundação para a Ciência e a Tecnologia (Portuguese Foundation for Science and Technology) within project "FCOMP-01-0124-FEDER-022701".

6. REFERENCES

[1] T. Alan and M. Peter. SPECTRa-T Final Report July 2008. 2008.

[2] C. Borgman. The conundrum of sharing research data. *Journal of the American Society for Information Science and Technology*, 63(6), 2012.

[3] Y. Gil, E. Deelman, M. Ellisman, T. Fahringer, G. Fox, D. Gannon, C. Goble, M. Livny, L. Moreau, and J. Myers. Examining the Challenges of Scientific Workflows. *Computer*, 40(12):24–32, Dec. 2007.

[4] R. Heery and M. Patel. Application profiles: mixing and matching metadata schemas. *Ariadne*, (25), 2000.

[5] L. Jahnke, A. Asher, and S. D. C. Keralis. *The Problem of Data*. Number pub154. Council on Library and Information Resources, 2012.

[6] S. Jones, S. Ross, and R. Ruusalepp. Data Audit Framework Methodology, 2009.

[7] P. Lord, A. Macdonald, L. Lyon, and D. Giaretta. From Data Deluge to Data Curation. In *eScience All Hands Meeting 2004*, pages 371–375, 2008.

[8] L. Martinez-Uribe and S. Macdonald. User engagement in research data curation. In *13th European Conference, ECDL 2009, Corfu, Greece, September 27 - October 2, 2009. Proceedings*, volume 5714 of *Lecture Notes in Computer Science*, pages 309–314. Springer, 2009.

[9] C. Palmer and M. Cragin. Data curation for the long tail of science: The case of environmental sciences. *Proceedings of the 3rd International Digital Curation Conference*, 2007.

[10] H. A. Piwowar, R. B. Day, and D. S. Fridsma. Sharing detailed research data is associated with increased citation rate. *PLoS ONE*, 2(3), 2007.

[11] J. Rocha da Silva, C. Ribeiro, and J. Correia Lopes. Managing multidisciplinary research data: Extending DSpace to enable long-term preservation of tabular datasets. In *iPres 2012 Conference*, pages 105–108, 2012.

[12] C. Strasser and P. Cruse. The DMPTool and DataUp: Helping Researchers Manage, Archive, and Share their Data. *Research Data Management Implementations Workshop*, 2013.

[13] A. Swan and S. Brown. The skills, role and career structure of data scientists and curators: An assessment of current practice and future needs. Report to the JISC. 2008.

[14] M. Witt and M. Giarlo. Databib: An Online Bibliography of Research Data Repositories. *ALA Annual Conference*, (Paper 2), 2012.

Building Institutional Capacity in Digital Preservation

Matt Schultz
Educopia Institute
Atlanta, Georgia 30309
matt.schultz@metaarchive.org

Mark Phillips
University of North Texas
Denton, Texas 76205
mark.phillips@unt.edu

Nick Krabbenhoeft
Educopia Institute
Atlanta, Georgia 30309
nick@metaarchive.org

Stephen Eisenhauer
University of North Texas
Denton, Texas 76205
stephen.eisenhauer@unt.edu

ABSTRACT

The *Chronicles in Preservation* project, being led by the Educopia Institute, is undertaking research to evaluate the degree to which several of the current digital preservation standards in use today (e.g., OAIS, TRAC, PREMIS, METS, etc.) can be applied to the diverse and at-risk content genre of digital newspapers. Institutions need guidance on incremental, skilled approaches and lightweight tools and resources if they are going to begin caring for such content in achievable and yet sustainable ways. The *Chronicles* project has researched, experimented, and begun advocating for a variety of skills, tools, and other resources that both embrace the current standards and seek to implement them in lightweight ways. They incorporate, apply and extend a number of existing as well as leading edge advancements such as BagIt, the DAITSS Description Service, UNT's PREMIS Event Service, and the NDSA Levels of Preservation.

Categories and Subject Descriptors

D.2.7 [**Distribution, Maintenance, and Enhancement**]: Extensibility; E.5 [**Files**]: Organization/structure; H.2.1 [**Logical Design**]: Data models; H.3.7 [**Digital Libraries**]: Collection, Standards

General Terms

Documentation, Design, Experimentation, Management, Performance, Standardization

Keywords

BagIt, DAITSS Description Service, Digital Newspapers, NDSA Levels of Preservation, PREMIS, PREMIS Event Service, Standards

1. INTRODUCTION

The *Chronicles in Preservation* project, being led by the Educopia Institute, has been contributing to the recent trend towards helping institutions take more manageable and incremental steps toward accomplishing their digital preservation. It has been doing so by researching institutional capacities for implementing existing standards (e.g., OAIS, TRAC, PREMIS, METS, etc.), and doing so in the context of one highly valued, yet at-risk set of digital assets—digital newspapers. The project has discovered that institutions need more lightweight approaches, less imposing data models, improved guidance, and non-sophisticated technologies in order to begin accomplishing their digital preservation and laying a foundation for more robust activities down the road. This paper will explain the trend toward incremental approaches; the research done in the *Chronicles* project that underscores the need for such approaches; and how the project is producing skills, technologies, and other resources to meet those needs.

2. BUILDING CAPACITY

Digital newspapers are a valuable, unique and at-risk set of scholarly assets. For more than a decade, stewards of historical newspaper holdings in the U.S. have been hard at work under the United States Newspaper Program (USNP) and the National Digital Newspaper Program (NDNP) to microfilm, catalog, digitize, archive and make accessible newspapers in the public domain. Under NDNP, the technical standards for digitizing this massive corpus of materials have achieved approval and uptake more broadly. Institutions seeking to digitize their newspaper holdings for long-term preservation now have a set of highly reputable and open standards to follow.

The NEH-funded *Chronicles in Preservation* project `http://www.metaarchive.org/neh`, is seeking to evaluate the degree to which the NDNP standards, and digital preservation standards more broadly (e.g., OAIS, TRAC, PREMIS, METS, etc.), can be applied to digital newspapers going forward, particularly in an environment where grant funds are becoming more scarce. The *Chronicles* partners all value the importance of following standards for achieving sound digital preservation but have first-hand knowledge that doing so can be costly. For that reason, the *Chronicles in Preservation* project is attempting to evaluate the current needs for preservation readiness of digital newspapers in all its wide diversity of forms (including born-digital and digitized), and identify the proper application of standards along a spectrum of achieving a minimum *essential* level of conformance up to a more robust *optimal* level of conformance. The hope being that stewards of digital newspaper collections can understand what they can achieve in the short-term with limited resources, and work their way up towards over the long-term with respect to existing standards.

3. IDENTIFYING SOLUTIONS

To make better sense of the current state of digital newspaper holdings, and the degree to which standards, and preservation oriented technologies have been applied toward their maintenance, the *Chronicles in Preservation* project has carried out a number of assessments, including:

1. a collections readiness assessment survey;
2. a sample data analysis; and
3. a focused set of interviews with digital newspaper stewards and curators (including the project partners, commercial publishers, state libraries, NDNP participants, as well as non-NDNP participants).

Each of these assessments have helped the project staff and partners gauge the gaps in resource availability for achieving various levels of conformance toward standards, and more importantly how best to improve and develop new skills, tools, and other resources that can help digital newspaper stewards to begin meeting various tiers toward preserving these valuable, unique and at-risk set of assets.

To begin with, a survey was formed that queried the project partners in four major areas, namely:

1. Collection & Repository Information: Partners were asked about the size and scope of their collections, formats, repository systems and other storage media in use, and whether they had ever been required to restore their collections under any scenarios of loss.
2. Collection Data Management: Partners were asked about their data management practices, including what sorts of object identifier schemes and file naming conventions were being used, how their newspaper data was structured, and the nature and extent of any metadata (particularly preservation metadata) that had been defined.
3. Preservation Assessment: Partners were asked about incidences of obsolescence or format migration or conversions for their digital newspaper files, if any, and what sorts of tools may have been used to manage such activities, as well as their perceived capacity for beginning to manage their digital newspapers from a more robust preservation perspective.
4. Ingest & Recovery: Partners were also asked about rates of collection growth, nature of changes and remediation, and any policies and practices that would have an impact on their ability to package and exchange their digital newspapers for a separate preservation system and what the parameters might be for recovering and rebuilding any preserved collections in the event of local loss.

It was found that institutions had a wide variety of local repository implementations and data management practices for their digital newspapers, had begun to do little more than routine backup for their content, and were not very far down the road toward applying preservation standards or technologies. It became very clear from the survey that digital newspaper stewards would require lightweight approaches for beginning to advance toward more standards-oriented practices for managing their digital newspaper collections.

Secondly, to observe first-hand the state of these digital newspapers, we proceeded to request sample newspaper data from each of our project partners Partners were asked to provide at least one full issue (up to 8 GB) worth of newspaper data for analysis. What emerged was that institutions had a range of different title/issue and sub-foldering schemes for their data, a variety of file-naming and object identifier schemes (often imposed by their repository/access systems), and varying amounts of descriptive, technical, administrative, and structural metadata. This made it clear that much work might be needed to apply some consistency across their collections for the purposes of packaging them for long-term preservation, and that this could prove to be a barrier for taking action in the short-term. Less imposing data models for preservation packaging were clearly in need.

Finally, effort was taken to reach out to stewards and curators of digital newspapers outside of the project to gain a broader perspective on the vast array of preservation challenges that may be facing such institutions. Interviews were arranged with a social media reporter for the Calgary Herald, a newsroom librarian at the Dallas Morning News, the State Librarian of the Wyoming State Library (Wyoming Newspaper Project), the University Librarian at UC Berkeley (California Newspaper Program), and the State Archivist at the Minnesota Historical Society to better understand how both born-digital and digitized news is being created, acquired, and managed in those contexts. In these interviews the urgency to get digital news under preservation quickly in the face of numerous institutional obstacles and barriers to partnerships was underscored. Institutions need help navigating existing standards, applying them in reasonable ways that respect their current capacities, and doing so with non-sophisticated technologies.

The *Chronicles in Preservation* project is working towards meeting this need by proposing, testing, and validating a combination of lightweight skilled approaches, technologies, and other resources to demonstrate how a diverse and complex set of digital assets like digital newspapers can be better curated in line with a tiered-spectrum of standards adoption and conformance. Below we talk about each of these skills, tools, and other resources. They include:

1. BagIt: Institutions need a beginning preservation data model in the absence of a consistent existing model;
2. Preservation Readiness Plans: Institutions need an incremental roadmap for improving curation and preservation packaging over time;
3. Simplified Preservation Metadata: Institutions need simpler (PREMIS) creation and management tools that can build off of data models like BagIt; and
4. Levels of Preservation Metadata Guidelines: Institutions need guidance on enhancing their data model and preservation metadata in incremental ways over

time. The NDSA Levels of Preservation are proving helpful in this area.

4. BAGIT DATA MODEL

In light of the information gathered in the previously mentioned survey and analysis of contributed test data, it became clear that there were a number of data models in use at the various partner institutions. Much digital newspaper content is being organized without a unifying data model that would allow institutions to make assertions and discuss characteristics of their underlying data in a consistent way. In order to resolve this challenge a decision to implement a data model utilizing the BagIt packaging specification was made [2]. The BagIt packaging specification has been used by a number of collaborative projects to package and share data between different technology and organizational platforms and was seen as an easy step towards a simple data model for the *Chronicles* project.

In order to implement the BagIt specification in the project the project team compiled a list of commonly used and maintained open-source BagIt tools, and documented them for project participants. In addition to the identification of these tools, we prepared a set of simple instructions outlining the tools and their use in the project. This documentation included installation information as well as a guide to the metadata fields (bag-info.txt). All of the partners reported success in making use of the simplified instructions. Making use of BagIt for their collections provided the partners with an opportunity to revisit their data structures, apply a simple packaging scheme for that data, and in many cases provided them with a previously non-existent layer of information (inventory and checksums) that could be elaborated on further (as will be discussed below). These simplified BagIt usage instructions will be made available for other institutions to use along with all of the project's code products at the conclusion of the project in April 2014.

5. PRESERVATION READINESS PLANS

In collaboration with the partner institutions, the project team also created a set of preservation readiness plans that established a number of lightweight preservation steps that could be applied to the partners' digital newspaper collections. These plans included contact information, roles and responsibilities for the collection, as well as scope of the collection in relation to the Chronicle in Preservation project. Additionally a series of goal statements followed by action plans for completing these goals were established for each partner.

These preservation readiness plans served as a starting point for conversation with collection owners to identify possible gaps in infrastructure, training, or tools at their institutions. A template example of these preservation readiness plans that other institutions can make use of will be made available along with all of the project's code products at the conclusion of the project in April 2014.

- Inventorying
- Checksums
- Format Identification

Inventorying files for the *Chronicles in Preservation* project involved partner institutions making explicit file-level inventories for content being used by the project. This inventory process aligns with the usage of the BagIt specification because the specification requires the creation of a manifest that defines the content of the valid bag.

Check summing of fixity information is another area of interest for the project participants. In sharing the readiness plans it became apparent that partner institutions varied widely in the tools used to generate fixity information and the use of that information for managing their digital newspaper collections. Identifying fixity as another area to focus was again complementary with the usage of the BagIt specification and data model for the project as the specification requires the inclusion of fixity information in the manifest in order to validate bags.

Finally format identification was identified as a goal in the preservation readiness plans. The readiness plans identified this as an area that would be more challenging for partners to implement locally than the previous two areas. A decision was made to build a set of identification services around the BagIt model that could be executed with limited overhead for the partner institutions. These services are described below.

6. SIMPLE PREMIS CREATION & USAGE

To simplify the process of performing format identification analysis over bagged collections of files, we leveraged the powerful DAITSS Format Description Service originally developed by FCLA (now FLVC) [1]. The service exists as a Ruby web application that can be run on a local machine, making it ideal for batch usage. We have developed a lightweight script, which when paired with the Format Description Service, can be used to analyze the entire contents of a bagged set of files and produce PREMIS records as output (stored within the bag itself). The script uses basic Unix commands to loop through the contents of a bag. Each file is sent to the Format Description Service (running on the local machine), and the resulting output is saved in a corresponding file inside a "premis" directory placed at the root of the bag. The output files are named and organized identically to the input files, with an ".xml" extension added at the end.

For more robust management and tracking of PREMIS data, we have also prepared the PREMIS Event Service software for general release in the near future. The PREMIS Event Service is a Django-based web application designed to manage and relate PREMIS records and related metadata in a database-driven system, originally built for internal use at UNT. The service provides a web-based user interface and REST API through which records can be fetched, queried, and stored in a way that allows for consistency and centrality throughout preservation systems [4]. With some light modifications, the scripting described above could be adapted so that the results of the Format Description Service is sent to the PREMIS Event service for storage. The project's code products will be made available and appropriately licensed for other institutions to make use of at the close of the project in April 2014.

7. NDSA LEVELS OF PRESERVATION

Finally, digital preservation standards such as OAIS and TRAC advocate for the application of a robust set of tools and practices to better accomplish long-term digital preservation, but these standards do not offer much practical guidance. The NDSA Levels of Preservation were formulated in response to this problem [3]. While they have been published only recently, the Levels have come to serve as a useful starter assessment resource for institutions. The *Chronicles in Preservation* project found the Levels especially useful for providing guidance on enhancing the proposed BagIt data model and preservation metadata in incremental ways over time.

The NDSA Innovation Working Group that is developing the Levels has suggested several methods for assessment including establishing a threshold level and providing an analysis for each stage and row of methods in the Levels. Because requirements for metadata can be found in all levels (even those outside of the metadata row), this assessment began by identifying all the suggested metadata requirements in the five categories:

1. Storage and Geographic Location: Important to retain metadata on accessible systems even in emergencies (Level 4)
2. File Fixity: Important to check or create fixity information for all objects on ingest (Level 1); virus check high-risk content (Level 2); check fixity at fixed intervals with logs, virus check all content (Level 3); check fixity in response to events (Level 4)
3. Information Security: Important to maintain logs of who performed what actions on objects (Level 3); audit logs (Level 4)
4. Metadata: Important to store object manifest separately (Level 1); store administrative and transformation metadata (Level 2); store standard administrative and technical metadata (Level 3); store preservation metadata
5. File Formats: Create an inventory of file formats (Level 2)

In the *Chronicles in Preservation* project, BagIt is a foundational tool (as described above). As mentioned, institutions often overlook creation of an object manifest. While the NDNP METS standards describe newspapers on an issue level, there is no requirement for a collection-level manifest. The BagIt specification includes a manifest of all objects in the bag with checksums. Creating and backing-up the manifest fulfills Level 1 Metadata. Bag validation utilities allow organizations to transfer bags and check fixity on ingest, in accordance with Level 1 File Fixity. The *Chronicles in Preservation* project also required the use of a bag profile to record administrative metadata for each collection such as the owning organization, contact information for the content's steward, the size of the bag, and a short description of the bag's contents, partially fulfilling Level 2 Metadata.

The Format Description Service mentioned above identifies file formats and creates corresponding PREMIS records once a collection or set of content has been "bagged". This process primarily accomplishes Level 2 File Formats requirements to inventory file formats, but wrapping the metadata in PREMIS also complements the administrative bag metadata in fulfilling Level 2 Metadata.

The final component of the Level 2 Metadata Requirements is logging transformative events that the organization performs on the objects over time. The creation of new derivative copies or the migration of master objects to new formats includes updating the metadata of the object. The PREMIS Event Service can provide ongoing monitoring of stored digital objects, allowing the organization to query changes in this metadata over time.

By utilizing the three tools above—BagIt, the Format Description Service, and the PREMIS Event Service—the *Chronicles in Preservation* project is able to automate the creation of nearly all metadata required below level 3.

8. CONCLUSION

Institutions engage digital preservation standards and methodologies with certain degrees of current capacity that determine what they can realistically accomplish in the short-term. There are legitimate trends in the community that are embracing incremental approaches. The *Chronicles in Preservation* project has underscored the need for such approaches and has sought to produce skills, tools, and other resources that embrace the current standards yet seek to implement them in lightweight ways—laying a foundation for more robust implementations over the long-term.

9. REFERENCES

[1] Florida Center for Library Automation. Daitss format description service, Apr. 2013.

[2] J. Kunze, C. D. Library, J. Littman, L. Madden, L. of Congress, and B. Vargas. The bagit file packaging format (v0.97). Internet-Draft draft-kunze-bagit-09, Internet Engineering Task Force, Apr. 2013.

[3] Library of Congress. Ndsa levels of digital preservation: Release candidate one. `http://blogs.loc.gov/digitalpreservation/2012/11/ndsa-levels-of-digital-preservation-release-candidate-one/`, Jan. 2013.

[4] M. Phillips, M. Schultz, and K. Nordstrom. Premis event service. In *Open Repositories 2011*, June 2011.

Adapting search user interfaces to web archives

David Cruz
Foundation for National Scientific Computing
Av. Brasil, 101
1700-066 Lisboa, Portugal
david.cruz@fccn.pt

Daniel Gomes
Foundation for National Scientific Computing
Av. Brasil, 101
1700-066 Lisboa, Portugal
daniel.gomes@fccn.pt

ABSTRACT

Despite the importance of web archives for the access to historical information published on the Internet, human interaction with web archives systems has not been thoroughly addressed. The web archive search user interface presented on this paper was derived from several rounds of development and usability testing over the Portuguese Web Archive search user interface (available at `archive.pt`). We present our findings gathered while adapting a typical web search user interface to the context of web archive search. We describe how we adapted a typical search user interface to address full-text and URL search over web-archived data, highlighting the unexpected problems detected during usability testing of our interface and current limitations for future work. The obtained results from usability testing showed that the average user satisfaction with our user interface was 70%. The obtained results from anonymous user satisfaction questionnaires yield a 84.3% score. We believe that our work can be applied to improve the quality of the services provided by other web archives.

1. INTRODUCTION

Most web users are not acquainted with web archives and accessing to archived web data is a significantly different user experience than accessing the live web [2, 5, 6]. Current users demand ready-to-use applications and are getting less tolerant to usability barriers. Although there is a significant number of web archives available [9], only preliminary research has been done about the design of user interfaces to gain access to temporal web data. The adoption of inadequate user interfaces to gain access to web archives jeopardizes the return of the investment made to preserve historical web content. User interfaces for web archives must be carefully designed and tested to respond to real-world user requirements and provide functional features specific to the exploitation of Web-archived content.

The Portuguese Web Archive (PWA) began in 2008. It is a public search service with over 1 131 million web files archived since 1996 that aims to preserve web content of interest to the Portuguese community (`archive.pt`). After witnessing the difficulties of our users, we increased our effort on improving their experience and satisfaction while using the PWA.

In this study, we share our experience of adapting a typical live-web search user interface to support web archive interaction. The lessons learned during this work would have been helpful to "kick start" our web archive in 2008. The developed code of our search system is freely available. We believe that our contributions will help other web archivists to improve the user experience and impact of their services.

2. METHODOLOGY AND RESULTS

The applied methodology followed an user-centered design approach [1]. The usability of our search interfaces was tested in collaboration with HCI experts from the Human Computer Interaction and Multimedia research group from the University of Lisbon. We tested several versions of our web archive user interface. Each new version triggered a new round of usability tests. Each testing round consisted of ten tasks presented to six users with no experience of using web archives. Each of the users executed the test individually in the presence of an usability expert. We recorded the screen, audio and participants' facial expressions for later analysis. Participants had to fill a questionnaire before the test to inform about their proficiencies on Internet usage and a post-test satisfaction questionnaire [4]. We finished each test with a debriefing session to further explore users' difficulties and clarify doubts in our observations. We obtained feedback from 21 users with distinct profiles. We analyzed the obtained results using a Likert scale from "1" (strongly unsatisfied) to "7" (strongly satisfied).

The obtained results from laboratory usability testing showed that the average user satisfaction increased from 3.6 (51.4%) on the first version of our interface to 4.9 (70%) on the last version. We also obtained results from anonymous user satisfaction questionnaires filled during dissemination events by users after freely trying the PWA, where we obtained a 5.9 (84.3%) score. This evaluation methodology occurred in an environment less controlled than our laboratory usability testing and used simpler questionnaires to avoid overloading users. On the other hand, the results were obtained from an usage environment closer to reality.

3. ANATOMY OF A WEB ARCHIVE SEARCH USER INTERFACE

Through usability testing on the first versions of the Portuguese Web Archive, we made two determinant observations. The first observation was that searching historical web content was an awkward concept to most web users. The existence of a website (web archive) that provides access to pages that are no longer available on their original websites is a perception that requires technical knowledge about the functioning of the Internet that is far beyond the skills of common web users. The second observation was that obliging the users to choose between a URL and full-

text search interfaces to gain access to web-archived content was ineffective and confusing to them.

The search user interface designed for the Portuguese Web Archive explores the users' familiarity with traditional search engines by offering similar layout and familiarity and enhance it with specific functions and contextual information required from web archives.

3.1 Search homepage

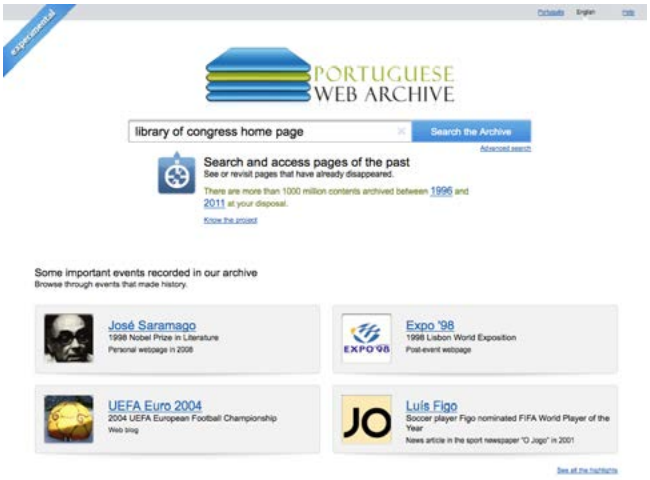


Figure 1: Interface for the home page of the web archive search user interface deployed on the Portuguese Web Archive.

Figure 1 presents the home page for the web archive search service. This homepage presents a search box without any temporal controls and some highlights of archived pages. The highlights are fundamental anchors that allow users to explore curated content, especially if the users never used a web archive before and have no clue about what they can find in it. The users observe examples that illustrate the type of information that they may find and progressively gain awareness about the potential of searching a web archive, thus reducing the cognitive effort of first-use. The publication of selected archived pages on the home page improved the overall user satisfaction with the service. Watching archived pages with historical value that have already disappeared triggered feelings of nostalgia which increased the positive perceptions, reflected through comments, about the provided service and general usefulness of web archives. Unlike most live-web search engines, our web archive search home page also includes a fat footer. The objective was to provide additional links to information that clarifies users about the context of web archiving and web archive search. For instance, links to: texts and videos about the project, a form to suggest sites to be archived, news or help.

We designed our interaction model to support both types of queries and shift the burden of detecting the query type to the system. Users only have to fill one search box and our system detects the type of query and presents the results in an interface tailored for that query type. When the query is composed exclusively by a URL, the corresponding version history results are returned. The results also include the versions from different URLs that are likely to reference the same content (e.g. www.site.pt, site.pt, site.pt/index.html). If the query is exclusively composed by text, the system returns full-text search results. If the query includes text and a URL, the system returns the full-text search results and suggests a link to the versions history of the URL. By doing so, the web archive interface becomes similar to live-web search engines, which users are already acquainted with, and guides them from familiar ground to the new context of searching historical web content.

3.2 Full-text search results

Figure 2 shows the user interface in full-text search mode. It is comprised by a typical search field for the query and a list of search results. But also, date input fields and datepickers to restrict the temporal interval of the queries. The two datepickers define lower and upper limits of the page archive dates to be searched. The results are shown on a results page similar to traditional live-web search engines. What differs is that we give greater emphasis to the archival date of each result. We tried several layouts and found that the position where users better recognized the dates was bellow the result title. Even so, some users ignored the unusual display of the date of archival within the search results. The interface for full-text search allows users to sort the results by relevance or archive date through the "sort:" operator or the advanced search interface. The advanced search also provides additional fields to allow more specific queries where users can restrict for: words, phrases, excluded words, file type, website or number of displayed results.

3.3 URL history search results

Searching for a URL or clicking on the "other dates" links on the full-text results page directs the users to the history view of that URL. The results are presented on a grid layout where each column group the several archived versions of a specific year, starting from the oldest year, supported by the Archive, up to the most recent year.

Each column then lists the available versions for that year, starting from the oldest. The users have an overall view of the versions available for a given URL. Clicking on the date link opens the correspondent version of the archived page. The grid layout approach was well understood by users. The versions from the current year are unavailable because the PWA only provides access to the archived pages one year after their archival so that the accesses to archived content do not concur with the original live-web sites (embargo policy). However, we display the current year column with a notice that explains the embargo policy to the users.

3.4 Reproduction of archived content

The interface currently in production that reproduces archived content presents a banner on the top of the archived page with the original URL and the archived date. Having an interface element always visible presents consistent hints that archived pages are different and behave differently from live webpages. However, we observed usability problems related to the reproduction of archived pages that deserve further research in future work. Users frequently lost perception about if they were navigating through archived content reproduced by a web archive or the live-web. One reason for this fact was that when the users scrolled-down the archived page, they lost visual contact with the top banner. On the other hand, the banner interfered with the layout of some

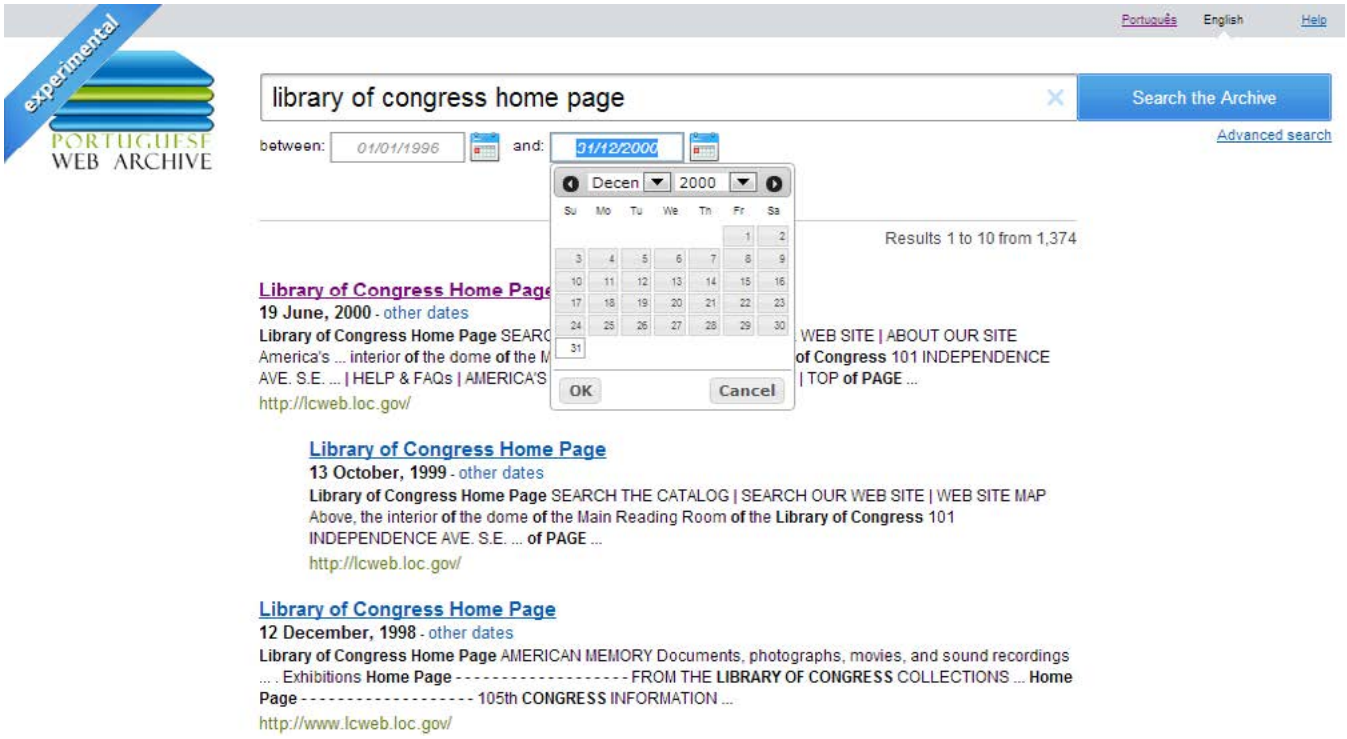


Figure 2: Interface for the web archive full-text search deployed on the Portuguese Web Archive.

archived pages by appearing on top of important content or links, for example in pages containing framesets or when the layout used absolute positioning in CSS.

Figure 3 presents the future interface design for reproducing an archived content. We took special care during the requirement analysis for this interface because it is unique to web archive search. Live-web search engines do not have to reproduce pages nor provide historical features, at most they provide a simple "cache" function that displays the textual content of the last version of an indexed pages without further concerns about maintaining its original layout. The presented design was derived through brainstorms using KJ-method [7] and several rounds of usability testing using low-fidelity paper prototypes [3] with a think-aloud protocol [8]. Notice that, unlike the previously presented interfaces, this new interface design has not yet been deployed to production on the PWA.

The archived page is reproduced in an internal frame so that the original layout is isolated and without interferences from the additional features on the page. We concluded that the interface for viewing archived pages should provide contextual information about the page (URL, date, help), features for sharing by e-mail and the main social networks (Twitter and Facebook), and for saving a copy of the archived page as image, PDF or compressed file. A sidebar enables the users to switch between versions of the archived page without having to return to the history page. To maximize the viewport devoted to the archived page, those contextual and navigational interfaces can be collapsed to a narrow bar above the archived page with the minimal information needed: The PWA logo that links to the Homepage, the URL, the date of the version presented and a button to expand the interface. Contrary to the old interface, the new one always show contextual information about the archived page, even if the archived pages are scrolled down by the users.

4. DATEPICKER TUNNING

The UI element that required the most tweaks was the datepicker. Standard datepickers are conceptually simple, only presenting a grid of the days of the month and left/right arrows to view previous/next months. However, web archives collect data that can span through decades. For example, the Portuguese Web Archive hold pages archived from 1996 to 2012. Thus, traversing this date range using a standard datepicker would require 203 clicks. After several design iterations, we concluded that a web archive datepicker should use drop-down lists to allow a quicker selection of month and year of the time span of the search (see figure 2).

We observed that for tasks with implicit days (e.g., "Movies released during June 2000"), users only specified the month and year but did not specify the day. Then, they either dismissed the datepicker by clicking outside (doing so closed it without saving the date) or became confused hesitating on how to proceed next. For the users, choosing the month was sufficient to communicate their temporal intent to the datepicker and got flustered because they had to do the extra work of choosing and clicking on a specific day. This unsatisfactory user interaction was overcome by adding a "OK" confirmation button and a "Cancel" button to dismiss the datepicker. With these buttons, users gained a strong visual anchor to decide unambiguously how to submit a new date or close the datepicker without any change to the current date. When the users click the "OK" button without selecting a specific day, the context of the datepicker determines its next state. If the user is defining the lower limit for

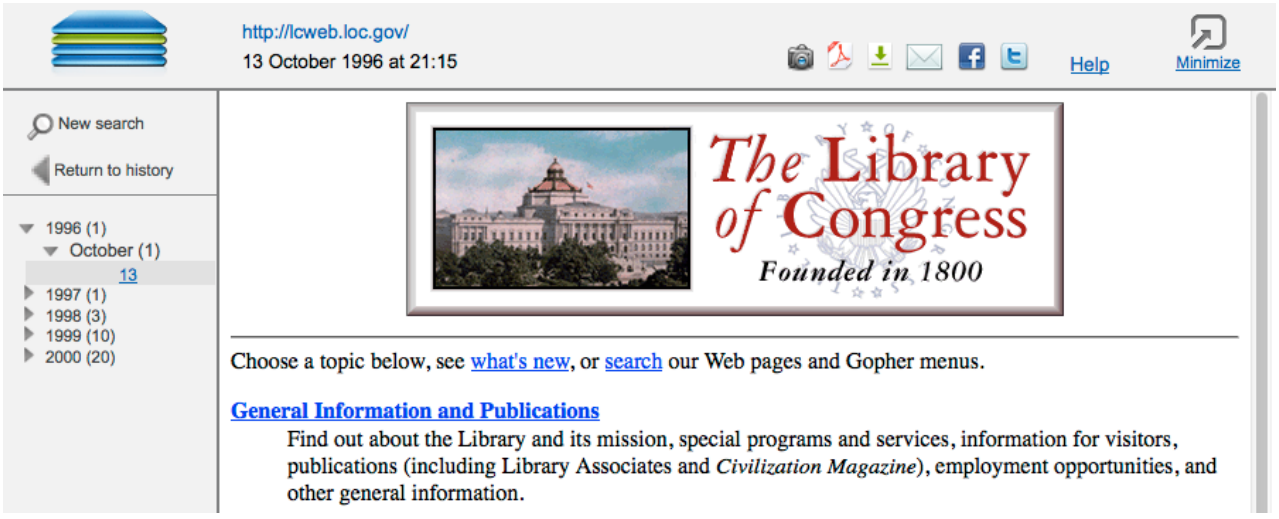


Figure 3: Interface design for reproducing archived pages.

the search interval through the left datepicker, the first day of the month is selected. If the user is defining the upper limit for the search interval through the right datepicker, the last day of the month is selected.

We also observed that some participants on the usability tests clicked first on the day before adjusting the month or year. The default behavior for the datepicker was to close immediately after the day was selected without leaving the opportunity for further adjustments. This user behavior depends on the date format the users were most familiar with. For example, for the date 24 December 1996, the users interacted with the datepicker according to their mental model of the date (day, month, year) and not to the visual organization of the information presented through the datepicker (month, year, day).

5. CONCLUSIONS

Search user interfaces for web archives must be similar to live-web search engines to facilitate the adoption of web archives by new users. This study presented the design and lessons learned while developing the Portuguese Web Archive (PWA) search user interface. The PWA follows the typical pattern of a live-web search user interface but enhances it with features to manipulate historical web content. Several aspects had to be carefully addressed in the design of the web archive search user interface such as how to handle different query types, how to present results that span across time or how to make users notice temporal information associated to archived pages. The results obtained from laboratory usability testing showed that the average user satisfaction increased from 51% on the first version of our interface to 70% on the last one that is currently in production at `archive.pt`.

Our main conclusions were that the home page for a web archive search service must present contextual information, such as examples of archived pages, that enable new users to gain awareness about what is and the potential of searching a web archive. Web archives should gracefully combine full-text with URL search. Users became unaware of their query misspellings, therefore web archives must provide query suggestion mechanisms. Standard datepickers are not adequate to be used in web archives and needed adjustments to be successfully applied as user interface elements to define the time scope of searches. By presenting the adjustments that we made to our interface and explaining their rationale, we expect to raise awareness about the importance of user interfaces to the success of web archives as useful services for modern societies. All the code and interface resources are freely available to be reused and improved at `code.google.com/p/pwa-technologies/`.

6. REFERENCES

[1] C. Abras, D. Maloney-Krichmar, and J. Preece. User-centered design. *Bainbridge, W. Encyclopedia of Human-Computer Interaction. Thousand Oaks: Sage Publications*, 37(4):445–56, 2004.

[2] P. Aravind, V. Arce, and P. Roessler. Qualitative Assessment of the Internet Archive's Wayback Machine. Technical report, University of California, Berkeley, May 2002.

[3] A. Coyette, S. Kieffer, and J. Vanderdonckt. Multi-fidelity prototyping of user interfaces. In *Human-Computer Interaction–INTERACT 2007*, pages 150–164. Springer, 2007.

[4] J. R. Lewis. IBM computer usability satisfaction questionnaires: psychometric evaluation and instructions for use. *Int. J. Hum.-Comput. Interact.*, 7:57–78, January 1995.

[5] J. Niu. Functionalities of web archives. *D-Lib Magazine*, 18(3/4), 2012.

[6] M. Ras and S. van Bussel. Web archiving user survey. Technical report, National Library of the Netherlands (Koninklijke Bibliotheek), 2007.

[7] J. Spool. The KJ-technique: A group process for establishing priorities. *User Interface Engineering*, May 2004.

[8] M. W. Van Someren, Y. F. Barnard, J. A. Sandberg, et al. *The think aloud method: A practical guide to modelling cognitive processes*. Academic Press London, 1994.

[9] Wikipedia. List of web archiving initiatives — wikipedia, the free encyclopedia, 2013. [Online; accessed 15-April-2013].

A Digital Archive of Monitoring Data

Fábio Costa
Faculty of Engineering, University of Porto
fabiopcosta@fe.up.pt

Gabriel David
INESC TEC, Faculty of Engineering, University of Porto
gtd@fe.up.pt

Álvaro Cunha
Faculty of Engineering, University of Porto
acunha@fe.up.pt

Rua Dr. Roberto Frias 4200-465 Porto, Portugal
+351-225081400

ABSTRACT
The change of status of data files from mere stepping stones to build other research products into publishable documents raises the question of how to organize data repositories appropriate for dissemination of such publications outside of the original research group. If the repository is to be used for the on-going research by the research group, it assumes the role of a digital archive. In this paper, a metadata model for the special case of projects relying on monitoring data is proposed and a prototype digital archive is described that has been built according to that model. This metadata is critical to preserve the context of production of the data at the organizational and technical levels and the meaning of each value. The digital archive offers several services for ingestion, visualization and dissemination that are essential for the effective adoption of the system. The method followed has been focus group work with a research group on structural health monitoring during the metadata specification phase, and an iterative development approach during the prototype construction phase of a digital archive for the same group.

Categories and Subject Descriptors
H.3.7 [**Digital Libraries**]: Collection, Dissemination, Systems issues

General Terms
Documentation, Design.

Keywords
scientific data repositories, experimental data streams, structural health monitoring.

1. INTRODUCTION
The investment and amount of effort put in setting up scientific experiments and collecting data from them fully justify that these data sets be properly preserved and eventually made available to the scientific community. This way, the research results may be cross-checked and used by other researchers for further investigations in what is being called e-science.

Due to the size of many of these data sets, it makes sense to organize their publication in data repositories, along with the required metadata to assert their meaning and authenticity. The metadata comprises contextual aspects on the entities involved and the purpose of data production, on authorship, on details about the scientific accuracy, on technical aspects of the digital support, and on integrity and preservation. The Core Scientific Metadata Model [5] covers most of these aspects. However, due to the diversity of scientific data, it is rather complex and may hinder the essential cooperation of the researchers in contributing the metadata elements.

The purpose of this paper[1] is to present a simplified model for monitoring data, which has been developed in dialogue with a team working on structural health monitoring; and a digital archive designed according to it, which is now being used as the research group's main data repository.

2. MAIN PROJECT GOALS
The data collection phase in research activity has mostly been considered a private concern of each project while the papers, reports and prototypes were the sole outcomes deserving to be published. Therefore, the collected data sets were organized in nonsystematic ways and, after being used, they were kept in the personal backups of the researchers and eventually discarded.

This understanding has been changing for several reasons. Some experiments are so expensive that it is not feasible to replicate them, as happens with high energy physics. The recording of natural phenomena, as in astronomy, is in many cases inherently unique. The advances in data acquisition systems led to the availability of huge data sets, in parallel with the capacity to process them. The development of the Internet turned the cooperation of research teams practical. All this has represented a strong push towards sharing not only the research results but also the data sets across the Internet. The creation of the Web itself has been a response to the need for cooperation in scientific research. Following the trend, several funding agencies adopted the policy of requiring the publication of the data sets produced within funded projects, which became research outcomes themselves [1]. The expertise required to properly design the experiments, decide and install the equipment and clean the data from defects and abnormal conditions in the acquisition is so high that the data sets can be seen as being authored by the researchers in charge of those tasks. Adding authorship to the data sets is a way to raise the personal responsibility of the researchers in properly taking care of the data sets, rewarding them by acknowledging their role in these scientific outcomes, and increasing the contemporary and future trust the data sets deserve.

Publishing means that the data sets will outspan the projects where they were born and even the research group. To make the

[1] Research supported by project "DYNAMO - Advanced Tools for Dynamic Structural Health Monitoring of Bridges and Special Structures", PTDC_ECM_109862_2009, funded by FCT/MCTES (PIDDAC) and FEDER through COMPETE/POFC. Gabriel David is co-financed by the North Portugal Regional Operational Programme (ON.2 – O Novo Norte), under the National Strategic Reference Framework (NSRF), through the European Regional Development Fund (ERDF).

data usable by researchers with no direct knowledge of the originating project circumstances, enough metadata must be added to the data sets. The metadata must fulfill two main roles: adding to the meaning of the data and improving the confidence on its accuracy and authenticity. So, it must include not only descriptions of the variables and measurement units, but also information on the process and equipment for data acquisition, on the researcher in charge of each phase of the project, and on the host institution, and details on the events that may affect the interpretation of the data, on the integrity of the data and on the processing the data may have been subject to.

Trying to encompass the diversity of situations that may occur in different research projects, leads to a complex metadata model. The context of the current paper is a research group working on structural health monitoring. The projects involve monitoring structures like bridges and other large Civil Engineering structures and their environmental and operating conditions. The monitoring is done through carefully designed and installed data acquisition units able to record, for instance, accelerations, temperatures, or wind speed. The data is collected in files every 30 minutes and sent by a data link to a computer of the research group. These raw data files are then pre-processed in order to clean possible malfunctioning situations and the cleaned files are also stored. One or more sophisticated algorithms are afterwards applied to the latter to calculate the evolution of relevant dynamic characteristics. The whole process may last for several years, resulting in a large number of similar and relatively simple files. The data in the results or processed files may be more complex, but it can be recalculated.

The importance of structural health monitoring is manifold. Keeping track of the behavior of bridges, dams, or large building under actual operating conditions of load, wind, earthquakes, etc. is important to study those structures and prevent incidents, to detect the effect of ageing and to help on repairing and compensating.

The main project goals are: (1) specify a metadata model; (2) design and build a digital archive, according to that model, able to store and organize the monitoring data as well as the processed results of the on-going projects; (3) improve data reliability through an integrated backup strategy; (4) create a Web interface able to browse and search the digital archive metadata and to visualize the data and download it; (5) set up a simple user management system and an access control policy; (6) automate the ingestion procedure of the data files into the digital archive.

Attaining these goals means a more reliable and systematic data life cycle, reduced researcher time on data management activities, support for data sharing in research cooperation, and a way to fulfill possible requirements of data publication. Furthermore, several important steps are taken towards preservation when insisting on collecting contextual and technical metadata and on organizing data in a systematic way.

3. METADATA MODEL

The metadata model is organized in three levels: Context, System and Data. There are two support classes related to the three levels, namely Person and Document. The Context package represents the information on the project itself and the hosting institution as well as the target structure being monitored. The project designs and installs specific data acquisition systems and chooses or develops specific software, both of which are described in the System package. Each data stream coming from an acquisition system is described in the Data package that also records the corresponding list of data files. The data files are organized in the file system.

According to the relevance given to authorship and good documentation, the support classes Person and Document are omnipresent in the model. It is possible to document the project and the structure, the data acquisition system, and the data sets with several types of documents, including technical descriptions, papers, pictures and diagrams. The documents have authors but, besides authorship, persons are associated to several components of the model, under different roles.

3.1 Context level

Contextual information (see Figure 1) is essential to know who and why has produced the data and under which circumstances.

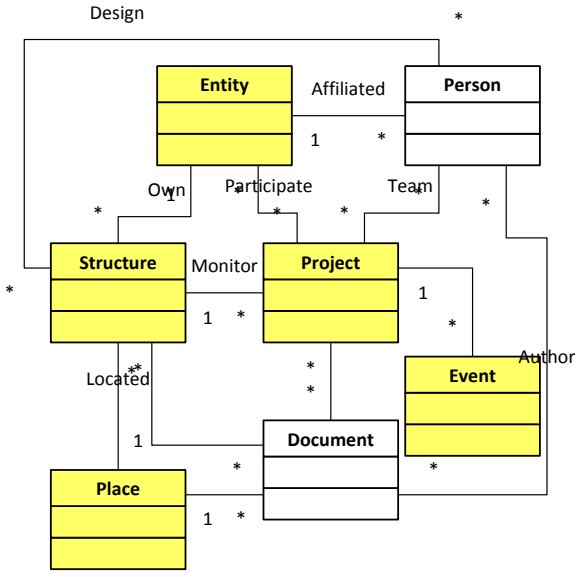


Figure 1. Contextual metadata.

All data sets are organized inside a project. So the central class is Project, including identification, type, life span and related entities. These entities (class Entity) may play different roles, like participating entity, funding agency, or owner of the monitored structure. The Project references a single Structure it monitors. The class Structure gathers information about the monitored object, like identification, building date, description, owner, location, and designers.

To improve on contextual information, the class Place records addresses of entities and geo-references structures and documents, in particular, pictures. There is also a class Event associated to Project that is meant to record any kind of event that may affect the monitored object or the data streams. Examples of events are earthquakes and strong winds, but also power shortages or maintenance actions on data acquisition systems. Events have an interval to enable setting a window on the data streams.

Structures and Projects may have associated documents of several types. Any kind of document may be associated but the recommended formats are PDF, PNG, JPEG and TIFF or any open document format, for preservation reasons. Besides the title, description and dates of creation and last update, technical details are recorded like the generating application, file type and size. It is possible to associate a place, especially meaningful for pictures. Documents relevant for structures are design summaries,

historical notices, and illustrative pictures. With respect to projects, the project proposal and a global diagram of the monitoring approach can be helpful. Each document has usually one or more authors, who are represented in the class Person.

Like documents, persons connect to the model in several points. Persons have identification and contacts and are affiliated to one entity. They are authors of documents and designers of the structures. And they group in teams for each project, with a certain role and during an interval. Persons belonging to teams and other designated collaborators have access credentials.

3.2 System level

The second level is about the technical system information (see Figure 2).

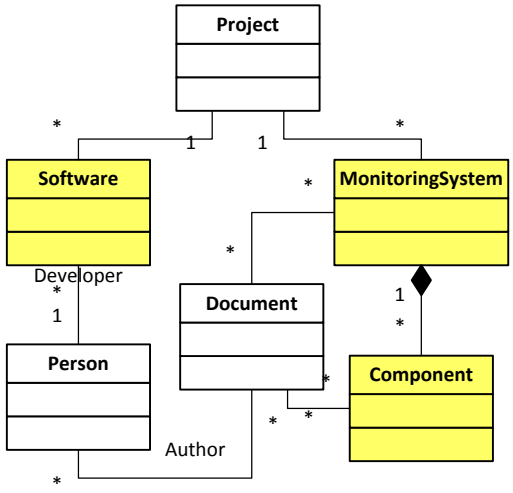


Figure 2. System metadata.

A project may use one or more data acquisition systems, simultaneously or in different periods. The main class is MonitoringSystem and it identifies a data acquisition system, its manufacturer and supplier, the period of operation and includes a description. A single data acquisition system usually possesses several transducers or data acquisition devices, with specific characteristics that are relevant to physically interpret the data streams. So, a Component class provides the details for each data source, like type, manufacturer, configuration, and positioning. Monitoring systems and components may have associated documents, like detailed installation diagrams, data sheets, or operating instructions.

Very often, the data acquisition process includes processing steps using commercial tools or specifically developed algorithms. The corresponding information, recorded at the system level in class Software, is the type, product, version, and the manufacturer or the developer, along with a description of its function.

3.3 Data level

The third level describes the data streams produced by the systems of the second level (see Figure 3). Each data stream is associated to a monitoring system and is represented by the Dataset class. It contains attributes describing the data stream, the acquisition method and the specific parameters used to obtain it. There is also a description of a possible processing step and a reference to the corresponding software. The intended results are summarized. A set of temporal attributes is also included like the period from one data file to the next, the number of files per day, and the sampling period and frequency. Some summary attributes include the time of the first and last reading in the data stream and the daily and global data volumes. The actual data location is recorded in the directory attribute. The types of foreseen data streams are: raw, for the data files as they are received from the acquisition system; pre-processed data streams correspond to a cleaned version of the data, after spurious values have been removed or errors have been fixed; and results data streams are those obtained by applying specific algorithms. The implicit genealogy of data streams is recorded in a many-to-many association.

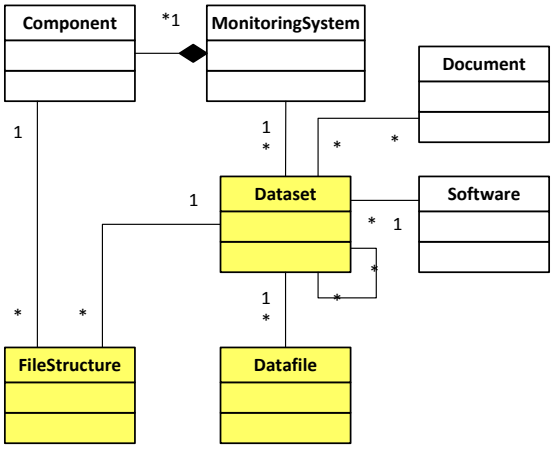


Figure 3. Data stream metadata.

A single data stream corresponds to a sequence of data files, all with the same record structure. The structure of a record is a sequence of columns. Each column is described by the class FileStructure. A column is associated to the channel of the component of the monitoring system producing that specific variable. The column has a number, two names (allowing for grouping similar columns), information on the type of variable and measurement unit, and a data type (integer, float, double, string). There is also an optional description. This information should be enough to understand and process the data files.

Finally, the Datafile class keeps track of the actual data files for each data stream. The main attributes are the filename, the file type, the file creation date, the start and end timestamps, the number of records, the file size and the compressed file size, a status (if the file is damaged) and a comment.

4. THE DIGITAL ARCHIVE

The metadata model of the previous section has been tested in the development of a digital archive for a research group in structural health monitoring[2]. The typical situation for projects in this area is to produce tens of thousands of datafiles.

4.1 Repository organization

The first problem on building a digital archive for these volumes is to establish an organization for the files. The decision has been to have a root directory for the digital archive, with a subdirectory for each project containing one directory for the documents associated to the project and one directory for each data stream. The latter is then hierarchically organized with directories for each year, each month and each day.

[2] ViBEST: http://paginas.fe.up.pt/vibest

4.2 Services in the prototype

The prototype that has been built is now in the first phase of use[3]. It implements the metadata model but in a certain sense is more than a data repository as it includes several services helping the research group to manage large amounts of data, use them in their day-to-day research and make them accessible to external researchers.

The technology used is the Postgres database management system, the Vaadin framework for Java Web applications [2], the Apache http server running on Ubuntu operating system and a few libraries for specific operations. The application has a Web interface automatically displaying pictures of each of the structures being monitored.

The generic information about each structure and its monitoring projects is publicly available. To go into more details about the monitoring systems and the data streams, authentication is required. So a simple user management module has been implemented, granting access rights at the project level.

In order to allow for manual as well as automatic ingestion of the data files, a background job has been created that periodically checks whether new data files have been received after the last checkpoint in each data stream and updates accordingly the records on the data files.

The design of the interface follows a compact style, trying to concentrate the most information on a single page. So, there is one page for the project and the corresponding structure, another for the monitoring system and a third one for the data stream (see Figure 4).

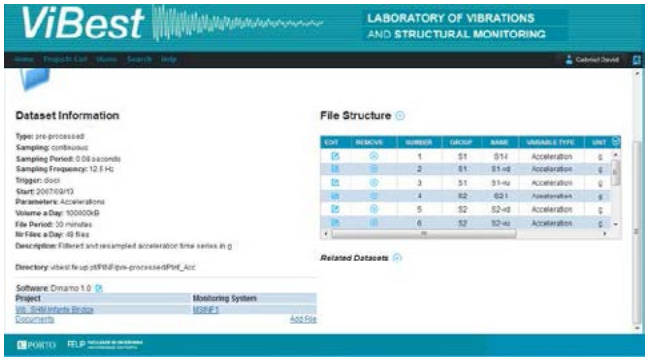


Figure 4. A data stream page.

However, due to the typical number of data files, these are only accessible through a selection form to choose an interval or an associated event. When arriving at a specific set of data files two options are given to the user: download them or visualize the data. The visualization option includes selecting which columns in the data files are to be included in a graph.

Two more aspects, related to dissemination, deserve mention. Due to the large number of files, downloading them one at a time is not feasible. So, a zipping facility has been prepared to combine all or, at least, chunks of the selected data files on a single zip file.

The second aspect is the addition of the OAI-PMH protocol [4] to enable the aggregation of the information in the digital archive by specialized repositories. A Dublin Core view on the metadata has thus been defined to support interoperability. This protocol works fine with the public information on projects and structures. However, the policy of the archive requires authentication for access to the second and third levels of metadata and to the data files. So, an extension to the OAI-PMH protocol has been prepared to allow authenticated users to keep using it at the level of data sets and data files.

At the same time that it improves the current research conditions, the prototype sets up the conditions for some preservation steps. The raw and pre-processed data files are zipped text files that will remain straightforwardly accessible. The metadata explaining the meaning and units of each variable, the sampling conditions, and the context of the experiment is collected in a relational database. The metadata is then preserved using the SIARD Suite [3] to convert the database into an XML representation.

5. CONCLUSIONS

The goals set up for the project have been achieved. In particular, the metadata model, although a bit demanding for the researchers asked to input the required information, proved to be enough to describe monitoring data for special Civil Engineering structures. As just few specific details of the structural health monitoring area have been used, the model is believed to be useful for general monitoring data projects.

With respect to the digital archive application, the size of the problem prevented the use of naïve approaches and forced some fine tuning of the http and application servers.

The main point still under analysis is related to the visualization of result data files not in a tabular format (scattered graphs, two or three dimensional matrices, etc.). Although it is possible to store these files in corresponding data streams, more work is needed in order to find an appropriate description for those data files. Probably, an XML representation will be chosen. A mechanism to visualize the diverse data formats needs to be devised.

6. REFERENCES

[1] European Commission. 2012. *Scientific data: open access to research results will boost Europe's innovation capacity*, press release IP-12-790. Brussels 2012-07-17. http://europa.eu/rapid/press-release_IP-12-790_en.htm?locale=en

[2] Grönroos, M. 2012. *Book of Vaadin*, 4th ed., pp. 466, Vaadin Ltd, Finland. https://vaadin.com/download/book-of-vaadin/vaadin-6/pdf/book-of-vaadin-pocket.pdf

[3] Heuscher, S., Järmann, S., Keller-Marxer, P. and Möhle, F., 2004. *Providing authentic long-term archival access to complex relational data*. In *Ensuring Long-Term Preservation and Adding Value to Scientific and Technical Data*. European Space Agency.

[4] Lagoze, C., Sompel, H., Nelson, M. and Warner, S. 2002. *The Open Archives Initiative Protocol for Metadata Harvesting, Protocol Version 2.0 of 2002-06-14*. Document Version 2008-12-07T20:42:00Z. OAI. http://www.openarchives.org/OAI/2.0/openarchivesprotocol.htm

[5] Sufi, S. and Mathews, B. 2004. *CCLRC Scientific Metadata Model : Version 2*, CCLRC Technical report DL-TR-2004-001, 2004. http://epubs.stfc.ac.uk/bitstream/485/csmdm.version-2.pdf

[3] Digital archive URL: http://vibest.fe.up.pt/shm

The Data-at-Risk Initiative: A Metadata Scheme for Documenting Data Rescue Activities

Anona C. Earls*, Erin Clary*, Jane Greenberg, Aaron Kirschenfeld, Angela P. Murillo, W. Davenport Robertson, Shea Swauger
University of North Carolina, School of Information and Library Science, Metadata Research Center
300A Manning Hall, Chapel Hill, NC 27599-3360
*aearls@email.unc.edu; *eclary@live.unc.edu

William L. Anderson
School of Information
University of Texas at Austin
1616 Guadalupe
Austin, Texas, 78701
band@praxis101.com

ABSTRACT
The Data-At-Risk and Rescue Initiative (DARI), an extension of the international CODATA Data-at-Risk Task Group (DARTG), is investigating how to best document data rescue efforts. This poster reports on a metadata-driven content analysis, and presents a metadata scheme for documenting data rescue. Twenty data rescue projects were reviewed for background context, and seven metadata schemes in the areas of preservation and data description were analyzed via a content analysis. Version 1.0, Data Rescue metadata core, consisting of 13 core elements, is presented, and future directions are noted.

Categories and Subject Descriptors
D.3.3 **[Language Constructs and Features]**, *Data types and structures*; E.2 **[DATA STORAGE REPRESENTATIONS]** O*bject representation.*

General Terms
Documentation, Design, Standardization

Keywords
Data Rescue, Metadata Schemes, Documentation, Endangered Data, Scientific Data.

1. INTRODUCTION
The Data-at-Risk and Rescue Initiative (DARI) is a project under the Committee on Data for Science and Technology (CODATA) Data at Risk Task Group (DARTG) [3, 8]. Initial DARI activities focused on the development of a prototype inventory to document the existence of valuable scientific data that are at risk of being lost to posterity [1, 2, 10]. At-risk data are data that are fragile or deteriorating, data that are lacking sufficient metadata, or data that are not in formats that permit full electronic access.

As work on the data-at-risk inventory progressed, the DARI team recognized the need to understand scientists' perception of at-risk data [9] and provide an online resource where scientists, data custodians and other individuals could contribute descriptions of data rescue activities. The goal of the work reported on in this poster has been to address this latter need, and to contribute to DARI's effort to extend the data-at-risk inventory to include descriptions of data rescue activities. Documenting successful data rescue missions illustrates to scientists that data can be saved and made available, and provides an important record of work that can aid with planning future data rescue activities. The work presented draws from successful data rescue efforts and work conducted in preservation and metadata communities.

2. BACKGROUND WORK
Data rescue has been an important human endeavor throughout history. Perhaps the most profound 20th century event was the 1966 Flood of the Arno River in Florence, which drew attention to preservation challenges in libraries, archives, and museums. The international community gathered to conserve and restore historical treasures in many collections, including the Institute and Museum of the History of Science, which is known to house historical scientific instruments and significant scientific collections, including the works of Galileo.

Digital technology has enabled new methods of data rescue. Several notable efforts include: the Astronomical Plate Collection and Preservation in China project, which is an effort to rescue, catalog, and eventually digitize astronomical plates from several observatories in China; the Royal Observatory of Belgium project, which seeks to digitize astronomical plates from the 20th century; and the Dominion Astrophysical Observatory project, which is focused on the digitization of materials from Canada's largest optical astronomical observatory. Related projects focus on the planet's changing climate and ecosystem. For example, the Botanic Garden and Botanical Museum Berlin-Dahlem rescue effort uses reBIND workflows to transform biodiversity data stored in outdated database management systems into well-documented, standardized formats. These and other data rescue efforts are important if data significant to the pool of scientific knowledge are to be preserved. However, brief descriptions on a web page or project page may not be sufficient to highlight these data rescue efforts, for scientists and other researchers to find these data, or for sharing approach outcomes on a global scale.

The DARI team is extending the data-at-risk inventory to include descriptions of completed and ongoing data rescue efforts. Over the last several months, DARI researchers have engaged in discussions and an exploration of what elements are essential to simply, yet thoroughly, describe data rescue efforts. The lead author of this paper has also contributed to this undertaking via her master's paper research, and she has focused on the development of a core metadata scheme for describing data rescue efforts. The remainder of this paper presents the DARI team's overall work, and the work conducted by Ms. Earls to develop a prototype metadata scheme to document data rescue efforts.

3. OBJECTIVES

The goal of the research presented here was to design a core, functional metadata scheme for the description and documentation of endangered scientific data rescue activities, and to apply metadata in accordance with this scheme to known data rescue activities within a digital content management environment.

4. RESEARCH QUESTIONS

1) What are the main descriptive characteristics of known data rescue projects?
2) What existing metadata standards can be applied to describing a data rescue project as a whole?
3) What metadata elements are essential for describing data rescue projects in particular?

5. RESEARCH METHODS

Scheme development was pursued using a mixed methods approach. First, 20 data rescue projects were reviewed for contextual background. Second, a content analysis was conducted to further examine seven metadata schemes in the areas of data description and preservation [6, 7]. The background review of existing data rescue projects included the identification of existing metadata used to describe or report on the effort, and a review of literature that reported on the effort [4, 7]. The content analysis compared schemes via a crosswalk to identify similarities and differences. Basic, core metadata elements became evident and form the basis for the Data Rescue metadata core, version 1.0.

Descriptive Elements	Archaeology Data Service Guidelines	Data-PASS	DOAP	Dublin Core, v. 1.1	DCMI-TERMS	Goddard Core	IMDI (ISLE Metadata Initiative)	RSLP Collection Description Schema
Title/Name	x	x	x	x	x	x	x	x
Description	x	x	x	x	x	x	x	x
Methods								
Notes		x						x
Creator/Author	x	x		x	x	x	x	
Sponsor								
Contributor	x			x	x			
Dates	x	x	x	x	x	x	x	
Geographic Location		x			x	x		x
Associated Resources	x	x			x	x		
URL/citation			x		x			
Subject Keywords	x	x		x	x	x		x
Unique ID		x		x	x	x	x	x

Table 1. Comparison of metadata elements across schemes.

6. RESULTS

The specific outcomes of this work to date include: 1) a prototype inventory for documenting data rescue activities that will serve a reference function similar to that provided by the descriptions of at-risk datasets, 2) version 1.0 Data Rescue metadata core for describing data rescue activities, and 3) a selected set of data rescue activities that are described in the inventory. The current scheme is heavily Dublin Core based, and future work will explore the value of advancing this work toward an endorsed Dublin Core Application Profile [5].

6.1 Metadata for data rescue, version 1.0

A proposed metadata scheme (**Table 2**) of thirteen elements has been developed. This scheme includes core elements for describing data rescue. The goal of the scheme is to facilitate consistent description of data rescue activities. The scheme forms the basis of an input template and has been through base-level testing. The scheme has been integrated into the DARI inventory, a publicly accessible metadata repository, developed via Omeka and located at http://ibiblio.org/data-at-risk/.

Element Name	Element Description
Title*	The title (and any alternatives) for the project.
Description	A brief summary of the main focus, goals, aims, and/or objectives of the project.
Methods	A brief summary of the approach, methods, techniques, and/or processes (including tools, software, etc.) being used for the data rescue.
Notes	Other details pertinent to the project, such as background information or project history.
Creator*	Individual(s) or organization(s) who initiated and have overseen the data rescue effort. May include contact information.
Sponsor	Individual(s) or organization(s) who have contributed financially or otherwise endorsed the project.
Contributor	Other individual(s) or organization(s) who have contributed to the project; for example, project partners/collaborators (physical or intellectual efforts), contributors of data/materials, etc.
Dates*	Dates indicating when the project was initiated and when the project was completed. May also include important milestones or other significant dates associated with the project.
Location	Location where the project was/is being carried out (if applicable).
Associated	Any other important projects or work (in particular, other data rescue initiatives) associated with this project, or upon which this project has been built.
URL	A link to the project website and/or online documentation of the project.
Keywords	Keywords indicating subject content of the project.
Project ID	A unique ID# assigned to the project by the repository (optional).

Table 2. DARI proposed metadata scheme for the description of data rescue activities. * Indicates a required element.

6.2 DARI data rescue description

The DARI team is at the early stage of testing the scheme's ability to represent a range of data rescue activities. A screen capture for one of the rescue projects is presented below in Figure 1. To date, we have tested two rescue projects thoroughly, and work will continue over the coming months.

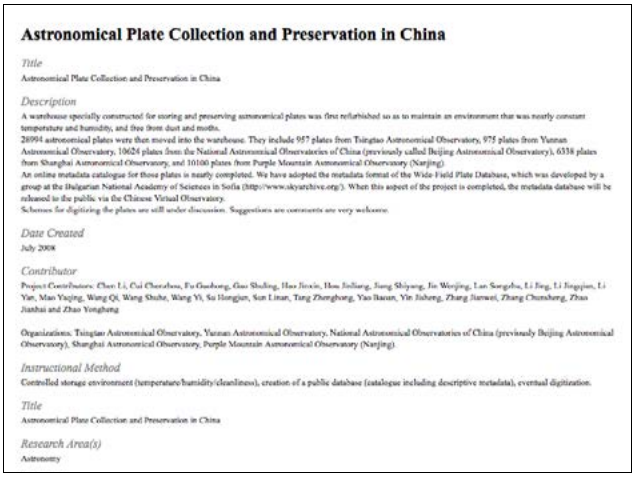


Figure 1. Screenshot, "Astronomical Plate Collection and Preservation in China," *Data-at-Risk Inventory*, accessed April 26, 2013, http://www.ibiblio.org/data-at-risk/items/show/94.

7. CONCLUSION

This paper reports on initial work to extend the DARI inventory and capture descriptions of data rescue activities. A core metadata scheme for documenting major identifying features of data rescue efforts, version 1.0 of the Data Rescue metadata scheme is presented, along with an example of a data rescue effort described with this scheme. A finalized scheme will serve to support knowledge and discovery of how endangered scientific data have been rescued in various cases. Future work will include further development of the inventory to support documentation of data rescue activities, and will seek to engage scientists, collection custodians, and other individuals in the documentation effort.

8. ACKNOWLEDGMENTS

We would like to acknowledge the support of CODATA and thank ibiblio for hosting the DARI Inventory.

9. REFERENCES

[1] Anderson, W., Faundeen, J., Greenberg, J., & Taylor, F. (2011). Metadata for data rescue and data at risk: ensuring long-term preservation and adding value to scientific and technical data. PV2011, 17 November 2011, Toulouse, France. Proceedings paper retrieved 2013-06-24 from http://hdl.handle.net/2152/20056. Conference presentation retrieved 2013-04-26 from http://www.slideshare.net/2ghouls/metadata-for-data-rescue-and-data-at-risk.

[2] Carver, N., Collins, K., Greenberg, J., Sinclair, J., Thompson, C., Veitch, M., & Anderson, W. (2011). Identifying endangered data: a case study supporting inventory design and implementation. ASIST 2011, October 9-13, 2011, New Orleans, LA, USA.

[3] Data-at-Risk Inventory. (n.d.). Retrieved 2013-04-26 from http://www.ibiblio.org/data-at-risk/

[4] Downs, R. R. (2009). Managing risks to scientific data. Prepared for presentation to the NYU/IBM Workshop on managing data risk: acquisition, processing, retention and governance, New York University, New York, NY April 24 2009. Retrieved 2013-04-26 from http://w4.stern.nyu.edu/emplibrary/Downs-ManagingRisksSciData.pdf.

[5] Dublin Core Application Profile, URL: http://dublincore.org/documents/profile-guidelines/.

[6] Dublin Core Metadata Initiative Metadata Terms, URL: http://dublincore.org/documents/dcmi-terms. Retrieved on 2013-04-26.

[7] Hodge, G., Templeton, C., & Allen, R. (2005). A metadata element set for project documentation. Science & Technology Libraries, 25:4, 5-23. doi: http://dx.doi.org/10.1300/J122v25n04_02

[8] International Council for Science: Committee on Data for Science and Technology. CODATA Data At Risk Task Group (DARTG). Retrieved 2013-04-26 from http://ils.unc.edu/~janeg/dartg/.

[9] Murillo, A. P., Carver, N., Greenberg, J., Robertson W. D., Thompson C. A., & Anderson, W. (2012). Data At Risk Initiative: Scientists' perceptions of endangered data and dat reuse. 23rd International CODATA Conference, 29-30 October 2012, Taipei, Taiwan.

[10] Nordling, L. (2010). Researchers launch hunt for endangere data. Nature, 468: doi:10.1038/468017a.

On Enhancing the FFMA Knowledge Base

Sergiu Gordea
AIT - Austrian Institute of Technology GmbH
Donau-City-Strasse 1
Vienna, Austria
sergiu.gordea@ait.ac.at

Roman Graf
AIT - Austrian Institute of Technology GmbH
Donau-City-Strasse 1
Vienna, Austria
roman.graf@ait.ac.at

ABSTRACT
Ensuring the long term access to digitized content is a major concern of digital libraries. The document migration and summarization are key activities employed reach this goal. The evaluation of preservation friendliness and making recommendations for long term preservation requires deep domain knowledge which is currently not available in any integrated knowledge base. In this paper we present an approach for enhancing the automatic aggregated knowledge on computer file formats. A clustering algorithm is employed to identify related file formats and to predict missing semantic associations between file formats and software tools. This is used to improve the discovery of software tools supporting the less popular file formats.

Categories and Subject Descriptors
H.3.7 [**Digital Libraries**]: System issues; H.3.3 [**Information Search and Retrieval**]: Clustering

Keywords
digital preservation, file format categorization, related file formats

1. INTRODUCTION
One important aspect of preservation planning is related to the file formats used for encoding the digital information. Currently, the information about the file formats is only partially available in domain specific knowledge bases and it is not appropriate formatted, accurate or complete in the open data repositories. The activities related to the preservation of digital content in libraries and archives are associated with high financial efforts, therefore the decisions about preservation planning must be taken by using rich, trusted, as complete as possible domain knowledge. There were significant efforts made in the last years within this research direction, but the systems built by now fail to effectively support preservation planning activities, mainly because they lack of a solid knowledge base (i.e. containing rich descriptions and contextual metadata related to available file formats)[9]. Typical preservation plans include migration of the content available in old file formats into formats that are preservation-friendly (e.g. well supported by standard hardware and software systems, appropriate for publishing on the web or on paper). One of the big challenges of preservation planning is to find the appropriate software tools that are available for executing the preservation plans, given the multitudediversity of available file formats, software tools and version incompatibilities. The migration pathways provided by PRONOM is limited, due to the fact that this information is manually collected by a relative small community. In contrast to this, the semantic web resources (i.e. DBpedia, Freebase) are supported by large communities, but they typically don't have a preservation related background. In consequence, these repositories contain rich descriptions of file formats, software tools and their vendors, but there is an extremely low coverage of the software to file format linking. This paper is a continuation of the work presented in [2] and it is intended to provide a solid knowledge base for the risk analysis module of the File Format Metadata Aggregator (FFMA) service [3]. The main contributions of this paper consist in employing clustering algorithms for identifying related file formats, making use of genre classifications and predicting missing semantic links between software tools and file formats.

The rest of the paper is structured as follows: Section 2 gives an overview on related work and concepts and Section 3 presents the domain specific issues related to the recommendation of digital preservation actions. Its subsections present the enhancements added to the FFMA knowledge base and the algorithms used within the proposed approach. Section 4 presents the setup, evaluation and the interpretation of the experimental results. Section 5 concludes the paper and gives outlook of the future work.

2. RELATED WORK
Preservation planning is one of the important topics in the digital preservation, which is one of the newest research fields of computer science. Within this context, tools like PLATO [4] were developed with the goal of creating preservation plans by scheduling different actions like identification, characterization and migration. It uses a cost based model for evaluating the effectiveness of document migrations and uses a knowledge base for storing facts about file formats and migration paths. Registries like P2 [9] and its successor LDS3 [8] concentrate on building a knowledge base using the linked data approach and computing the preservation risks for individual file formats. Similar to our approach these systems integrate information collected from PRONOM and DBpedia, but they do not compute any enrichments, classifications and do not predict missing semantic links. The unified digital format registry project (UDFR) developed a platform based on semantic web technologies, which allows editing descriptions for file formats that were imported from PRONOM and MIME media types repositories [1]. In extensions to simple metadata aggregation, the

approach presented in this paper uses artificial intelligence technologies for enrichment and reasoning on the formats descriptions. For inferring explicit knowledge on related file formats we employed the well known algorithm: Density-Based Spatial Clustering of Applications with Noise (DBSCAN) [6]. Furthermore text based information retrieval models are use for computing similarities between natural language descriptions [7].

3. APPLICATION DOMAIN AND RELATED ISSUES

The knowledge based recommendation technology is the natural solution chosen for implementing tools supporting the preservation planning activities. They typically make use of expert rules and have the goal to analyze the compatibility of the content repositories with the state of the art and future technical infrastructures. The risks of not being able to archive, render or publish digital objects with modern tools are estimated. In the following we present a simplified representation of the digital preservation recommendation problem by illustrating the core of the recommendation algorithms:

```
IF
    NotPreservationFriendly(in: Format A)
THAN
    FindPreservationFriendly(in: Format A, out: Format B)
    FindMigrationSoftware(in: Format A, in: Format B, out: Software S)
RECOMMEND
    MigrateContent(in: Format A, in: Format B, in: Software S,
                   in: Configuration C, in: File NPF, out File PF)
```

where the type of the input and output variables belong to: *Format* - file format used for encoding content, *File* - a digital file storing multimedia content, *Software* - software tool used for processing a given file and, the *Configuration* used by the software tools in data migration processes. Within the pseudo code displayed above one can identify the key research questions that need to be solved by digital preservation recommender system:

Computation of preservation friendliness. The preservation friendliness of a given file format can be estimated by analyzing its complete description. This depends on the type of the content (i.e. text, image, audio, video), institutional context (e.g. archiving vs. web publishing), being open or standardized format, being supported by major vendors, rendering and processing with open source software, etc. Advances on this research topic are presented in [3].

Identification of migration Software. Whenever the digital content is packaged in an obsoleted or inappropriate file format, it is recommended to migrate it to a new representation (i.e. encoding) that is compatible with the modern communication technologies and processing/rendering software. As this information is not explicitly available, either in (open) domain specific knowledge base nor in semantic web. We aim at discovering this important information by using two heuristics: a) A software that is able to process two different file formats is able to convert between the two encodings (e.g. typically accessible through "Save as.." action) b) Each software is meant to process a group or related or equivalent file formats (i.e. document processors, graphic software, multimedia software, etc.). Our efforts for automatic clustering of the related file formats are presented in Section 3.2.

Preparing migration configuration. The conversion of the content from one encoding into another one requires provision of encoding specific and software specific parameters. This is achieved by evaluating the software tools and experimenting with them for ensuring the required quality of the conversion. This research topic is addressed by the work carried out in projects like Planets [4] and SCAPE [5]. In the current paper we focus our attention on the second research issues and we aim at identifying candidate software tools that are able to open specific file formats. The proposed approach uses the genre classifications and the free text descriptions to discover similarities between file formats and to infer predictions on matching software products. The currently used algorithms are not able to provide a high level of confidence, since the software and the file format versions are not taken in account. This is due to the fact that the version information is not available in linked open data repositories, except for a very few items.

3.1 Enhancing the FFMA Knowledge Base

A detailed analysis of the content and the size of the FFMA knowledge base was presented in [2]. It contains rich descriptions of about 594 file formats, 3719 software tools and 63 vendors aggregated from PRONOM, Freebase and DBpedia repositories. Despite of the richness of individual item descriptions, one of the weaknesses of the FFMA knowledge base is the low coverage of file format to software tools linking as presented in Figure 1. This histogram shows the distribution and the coverage of the file format to software relationships in the aggregated database. There are 154 file formats for which no software is known and there are 474 software tools for which no more that 3 supported file formats are known. From digital preservation point of view,

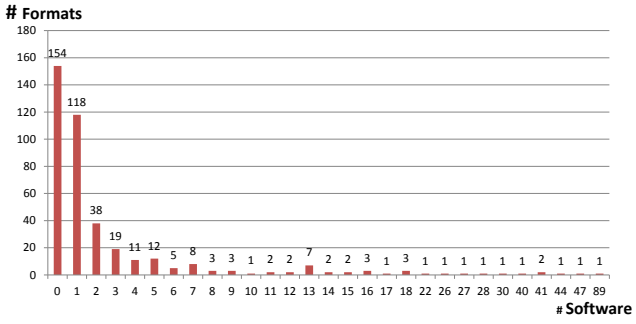


Figure 1: Histogram of software tool support for file formats

it is relevant how well a file format is supported, on how many platforms and how many software tools may render or edit it. In practice, the software tools are able to open more (related) file formats with different version (e.g. the most popular image file formats can be opened by the most of the image processing tools). By using a *linking through clustering approach* we aim at discovering important knowledge used within the preservation planning activities. For many software tools, there only a part of the list with the compatible file formats linked in the knowledge base, but there is a good chance that the tool is able to process additional similar file formats. For example, knowing that an graphic tool is able to process JPEG2000 files, there is a great chance that this tool will be able to process related file formats, like regular JPEG, Bitmap or TIFF. By using this reasoning, we aim at enhancing the digital preservation recommenders and enabling diagnosis in case that no migration solutions are provided. In this case, a set of candidate

software is generated including tools supporting related file formats (e.g. having similar genre and similar textual descriptions). External resources (e.g. homepages) might be manually checked to identify if one of the candidate tools is able to perform the conversion and to improve the recommender's knowledge base.

3.2 Related File Formats

For computing the related file formats clusters we use a variant of the most representative clustering algorithms, namely DBSCAN [6]. The ideas behind this algorithm is that the points within the cluster are mutually density-connected, which means in our case, that Format A and Format B belong to the same cluster in the case that each of the formats indicates the other one as being a neighbour. The definition of the algorithm is generalized and it is abstracted from the computation of neighbourhoods (i.e. distance between points in vector space).

The proposed algorithm uses textual information to compute distances between file format descriptions [7]:

$$dist(Q,T) = 1 - sim(Q,T) = 1 - \sum_{t \in Q,T} tf \cdot ln\frac{N}{df}, \qquad (1)$$

Where $dist(Q,T)$ represents the distance between query format description Q and the target format description T, which is the inverse function of the similarity between the formats $sim(Q,T)$. t stands for the terms found in both descriptions, N for the total number of format descriptions, while tf represents the term frequency within the target format description and the df represents the document frequency, respectively (i.e. in how many format descriptions the term t occurs). In the experimental evaluation we make use of the implementation provided through the "MoreLikeThis" handler available in Solr [1].

4. EVALUATION

The experimental evaluation was carried out by using the FFMA knowledge base and the genre classification of file formats available in Wikipedia. The goal of this evaluation was to show that the textual descriptions aggregated from linked data can be used to identify similar file formats. Furthermore, we evaluate the tool support on cluster level which provides input for enhancing the migration pathway generation.

4.1 Identification of related file formats

The identification of the related file formats is performed by using the algorithm described in the previous section. The results of the clustering is depicted in Figure 2 showing distribution of the file formats over the 29 clusters identified by our algorithm containing at least 5 members. The centroids of individual clusters are represented on the X axis, while the Y axis represents the amount of formats that belong to the given cluster. The largest clusters are represented by the following centroids: *doc*, *ace* and *dwg* with a member count of 70, 35 and 34 respectively. The *doc* labeled cluster contains the textual documents, *ace* stands for the archiving formats cluster and *dwg* (DraWinG) for standard raster formats. Clusters calculation evaluated 51 clusters with the

[1] `http://wiki.apache.org/solr/MoreLikeThis`

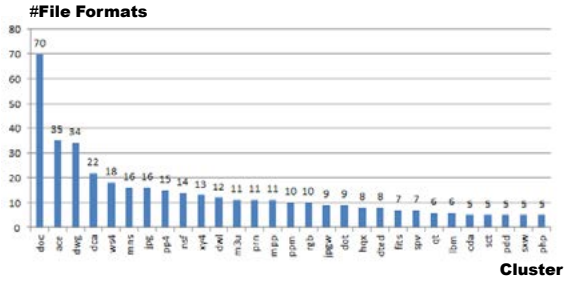


Figure 2: Distribution of file formats in clusters

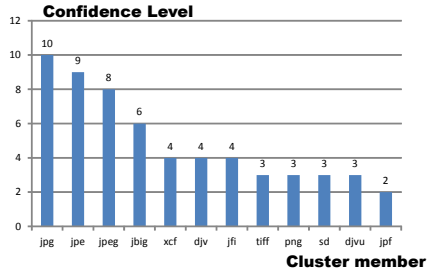


Figure 3: File formats in JPG cluster

nodes count in the range from 5 for *php* and 70 for *doc* cluster. Each cluster must have at least 5 members, otherwise the formats were considered as being outliers. Figure 3 presents the members of the *jpg*, most of them being well known raster graphics formats. The gain of the clustering

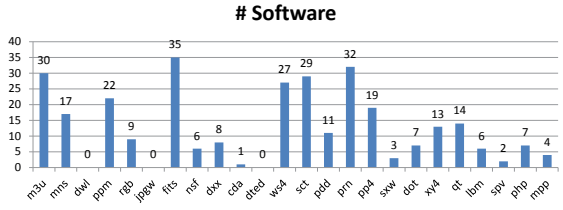


Figure 4: Software support for clusters.

consists in the identification of the software tools that are supporting several of the formats within the cluster. Figure 4 presents the association of the software support for the less supported file format clusters, indicating that most clusters have more than 5 tools associated. For a small part of the clusters there are still no or very few tools known in the database as being able to process the associated file formats (6 clusters supported by up to three software tools). Archiving formats, text processing and image file formats clusters with strong tool support (about 100 tools or more) are presented in Figure 5. In conclusion, the application specific and not standardized formats are supported by a lower number of software tools according to the current version of the knowledge base.

4.2 Classification of related file formats

An existing categorization of file format types was used to verify the hypothesis that the formats with similar textual descriptions are related to each other (i.e. using alternative representations or encodings of the same type of data, allowing data conversion from one format to the other, etc.). The *List of file formats* available in Wikipedia presents the assignment of file format extensions to their types, which

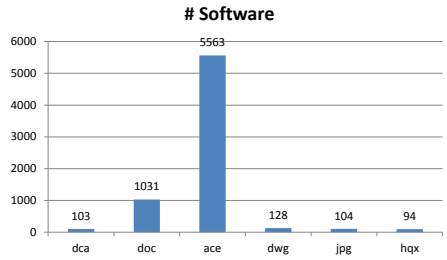


Figure 5: Clusters with strong software support.

Measure	Value
Number of file format clusters	30
Avg. file formats per cluster	13.6
Avg. format types per cluster	2.77
Avg. file formats of dominant type	21.09%
Avg. not classified file formats per cluster	64.26%
Avg. clusters with less than 2 categories	73.3%

Table 1: Statistics regarding the distribution of format types in related file formats (clusters)

are organized within a hierarchical structure [2]. In the first instance we used the Open Refine [3] tool for transforming the html representation of the categorization hierarchy to an appropriate taxonomy using the SKOS format. The later is available for download[4] from the FFMA server. The hierarchical taxonomy has the advantage of grouping categories of file formats that share certain commonalities (e.g. Raster Graphics and Vector Graphics are different possibilities of encoding Graphics content). The statistics on the type classifications of file formats over all clusters is presented in Table 1). There was accounted an average of 13.6 members per cluster and an average of 2.77 assigned format types (see also Figure 2 for cluster size distribution). The results presented here are highly influenced by the lack of categorization information, for 64% of the file extensions (available in the FFMA knowledge base) no file type assignment was found in the Wikipedia article. Under these circumstances about 21% of formats belonged to the *dominant* category and less than 15% was assigned to other categories. **Discussion.** The preliminary experimental results presented within this paper demonstrate the feasibility of the proposed approach. Anyway, no fine tuning of the clustering algorithm was performed, and no adjustments of the user generated taxonomy of file format types was made. Even so, the statistical analysis of the file formats presented in the Table 1 confirm our hypothesis that related file formats can be automatically identified using their descriptions (i.e. the average format types per cluster is 2.77, and 73% of the clusters have at most 2 categories). Still, this is not a strong evidence given the high rate of not categorized file formats. In time, we expect that more categorizations become available and the rate of formats of the dominant type to be significantly increased, even if the diversification of the format types per cluster might increase slightly. As future work we plan to significantly increase the rate of file format categorizations by taking in account more information sources like DBPedia genre, FileInfo classification [5], Yago formats [6], which will require spending significant efforts on ontology mapping purposes.

5. CONCLUSIONS

In this paper we present the enhancements added to the knowledge base of the file format metadata aggregator service. Artificial intelligence techniques are employed for identification of related file formats and to discover additional software tools that might be able to perform content migration between these formats. The preliminary evaluation demonstrates the feasibility of identifying similar formats basing on the textual descriptions acquired from the linked open data repositories. As future work we plan to use additional knowledge sources (e.g. vendor's web sites, further domain specific knowledge bases) for extending the knowledge related to the software tools, vendors and their relationship to the existing file formats.

6. REFERENCES

[1] U. C. Center. Unified digital format registry (udfr) - final report. Technical Report 2012-07-02, California Digital Library, University of California, 2012.

[2] R. Graf and S. Gordea. Aggregating a knowledge base of file formats from linked open data. In *iPress 2012*, pages 293–294, 2012.

[3] R. Graf and S. Gordea. A risk analysis of file formats for preservation planning. In *iPress 2013*, page to appear, 2013.

[4] H. Kulovits, C. Becker, M. Kraxner, F. Motlik, K. Stadler, and A. Rauber. Plato: A preservation planning tool integrating preservation action services. *LNCS - Research and Advanced Technology for Digital Libraries*, 5173:413–414, 2008.

[5] R. K. T. R. E. S. P. T. Orit Edelstein, Michael Factor. Evolving domains, problems and solutions for long term digital preservation. *iPRES 2011 - 8th International Converence on Preservation of Digital Objects*, 2011.

[6] J. Sander, M. Ester, H.-P. Kriegel, and X. Xu. Density-based clustering in spatial databases: The algorithm gdbscan and its applications. *Data Mining and Knowledge Discovery*, 2:169–194, 1998. 10.1023/A:1009745219419.

[7] A. Singhal. Modern information retrieval: A brief overview. *IEEE Data Eng. Bull.*, 24(4):35–43, 2001.

[8] D. Tarrant and L. Carr. Lds3: applying digital preservation principals to linked data systems. In *Ninth International Conference on Digital Preservation (iPres2012)*, 2012.

[9] D. Tarrant, S. Hitchcock, and L. Carr. Where the semantic web and web 2.0 meet format risk management: P2 registry. In *iPres2009: The Sixth International Conference on Preservation of Digital Objects*, June 2009. Event Dates: October 5th and 6th, 2009.

[2] see `http://en.wikipedia.org/wiki/List_of_file_formats`

[3] see `http://blog.semantic-web.at/2011/02/17/transforming-spreadsheets-into-skos-with-google-refine/`

[4] `http://ffma.ait.ac.at/taxonomies/FileFormatTypes`

[5] `http://www.fileinfo.com/`

[6] `http://dbpedia.org/class/yago/Format106636806`

A new data model for digital preservation and digital archiving for the French Administration: VITAM model on NoSQL technologies

Frédéric BREGIER
Ministry of Culture and communication, General Secretary, IT department
Rue du Fort de Saint-Cyr, Montigny-le-Bretonneux
78182 Saint-Quentin-en-Yvelines Cedex.
France / +33 1 30 85 67 49
frederic.bregier@culture.gouv.fr

Frédéric DEGUILHEN
Ministry of Foreign Affairs, IT department
3, rue Suzanne Masson
93126 LA COURNEUVE Cedex.
France
frederic.deguilhen@diplomatie.gouv.fr

Nathalie MORIN
Ministry of Defence, General Secretary for the administration, Memory, Heritage and Archives department, Archives and libraries policies office
14, rue Saint-Dominique
75700 Paris 07 SP, France
+33 1 44 42 12 35
nathalie.morin@intradef.gouv.fr

Marie LAPERDRIX
Ministry of Culture and communication, National Archives
59, rue Guynemer 90001
93 383 Pierrefitte-sur-Seine Cedex.
France / +33 7 86 55 17 12
marie.laperdrix@culture.gouv.fr

Lourdes FUENTES-HASHIMOTO
Ministry of Foreign Affairs, Archives direction
3, rue Suzanne Masson
93126 LA COURNEUVE Cedex. France

Edouard VASSEUR
Ministry of Defence, General Secretary for the administration, Memory, Heritage and Archives department, Archives and libraries policies office
14, rue Saint-Dominique
75700 Paris 07 SP, France
+ 33 1 42 19 71 41
edouard.vasseur@intradef.gouv.fr

Thomas VAN DE WALLE
Ministry of Culture and communication, National Archives
59, rue Guynemer 90001
93 383 Pierrefitte-sur-Seine Cedex.
France / +33 1 64 31 74 75
thomas.van-de-walle@culture.gouv.fr

ABSTRACT

The three ministries in charge of public digital archiving in France (Culture, Defence and Foreign Affairs) decided to build a specific system in order to preserve their digital information. The main challenge is the management of all the data and metadata produced by the French State which could be linked to Big data technologies. Since February 2013, these three ministries have done a large experiment (a proof of concept) based on NoSQL technologies, which ended in June 2013. In this paper, we describe our IT approach of this archivistic problem, our new data model and the results of this inter-ministerial study.

Categories and Subject Descriptors

C. Computer Systems Organization / C.4 PERFORMANCE OF SYSTEMS (Design studies, Fault tolerance, Modeling techniques, Reliability, availability, and serviceability)
D. Software / D.2 SOFTWARE ENGINEERING / D.2.10 Design (Representation)
E. Data / E.1 DATA STRUCTURES (Distributed data structures, Graphs and networks, Trees, Record)
G. Mathematics of Computing / G.2 DISCRETE MATHEMATICS / G.2.2 Graph Theory (Graph algorithms, Trees)
H. Information Systems / H.2 DATABASE MANAGEMENT / H.2.1 Logical Design (Data models)
H. Information Systems / H.2 DATABASE MANAGEMENT / H.2.4 Systems (Concurrency, Distributed databases, Query processing, Textual databases)
H. Information Systems / H.3 INFORMATION STORAGE AND RETRIEVAL / H.3.1 Content Analysis and Indexing (Indexing methods)
H. Information Systems / H.3 INFORMATION STORAGE AND RETRIEVAL / H.3.6 Library Automation (Large text archives)
H. Information Systems / H.3 INFORMATION STORAGE AND RETRIEVAL / H.3.7 Digital Libraries (Collection)

General Terms

Management, Measurement, Performance, Design, Economics, Reliability, Experimentation, Security, Human Factors, Standardization.

Keywords

Digital archiving, NoSQL, metadata.

1. VITAM, A JOINT PROJECT BETWEEN THREE MAJOR ARCHIVAL INSTITUTIONS IN FRANCE

The National Archives of France are in charge of archiving the documents produced by the French administration and government with the exception of two independent ministries: Defence and Foreign Affairs. The National Archives have decided to rethink their methods to collect, arrange, describe and preserve digital archives and to update their digital repository, CONSTANCE, which has been developed in the 1980's. There is an urgent need to build a new system in order to be able to meet the expectations of today's administration: adopting a "mass-production" approach has become a priority because of the exponential growth of digital information. CONSTANCE was set up at a time when the use of technology and technologies themselves were very different.

Thus, the National Archives launched a new project in 2011 called VITAM (the name of this project refers to the latin phrase *ad vitam aeternam*). As the National Archives are a department of the Ministry of Culture, they work closely with its IT Department; this collaboration is essential to build a solid model. Therefore, VITAM was included, as a strategic project, in the Ministry of Culture's IT outline plan in December 2011. From that date, the Ministry of Foreign Affairs and the Ministry of Defence, which are the only autonomous ministries allowed to keep their own historical archives, have joined the project. The three main archival institutions in charge of archiving the information produced by the French State are united around a common goal.

Controlling metadata: how to plan an intelligent access to digital information over the time?

VITAM's philosophy is directly in line with the legacy of CONSTANCE: simplicity, neutrality, durability, integrity. VITAM's functional model is based on the OAIS model [1] and also integrates records management standards (ICA-Req and MoReq [2]) in order to adapt the system to the needs of the French administration. The OAIS framework and vocabulary has been adapted for that purpose [3]. One of the major challenges of this new project is the description of digital archives and the capability to make requests in the new system over the time. In fact, the development of the information society has created facilities for copying, deleting, and editing documents, information and data produced by the public administration. However, data, as of paper archives, should be stored in specific conditions of integrity, security and authenticity. To meet these needs, it is necessary to assign to each given document or digital information many descriptive, archival and technical metadata.

Metadata has often much more value than data, information or original documents. They give meanings and make the archives intelligible. Moreover, in the context of increasing information sharing between different services, one must be able to hold this business details correctly. This fine archival description could not take place in the paper world due to lack of human resources sufficient to handle the mass of paper archives. However, information technology can multiply our processing capacity and allows us to consider keeping all these traces of digital information and ensuring their authenticity, integrity and intelligibility *ad vitam*. We have considered several solutions and we have built a specific model for metadata based on the National Archives experience and based on national and international standards.

Firstly, we will describe our IT approach and more precisely the use of a Big Data model to describe digital archives. Then, we will explain our model to describe and process metadata. Finally, we will present our experimental approach as a result of the Proof of concept we did for the IT director of the French State.

2. "BIG DATA" MODEL TO REPRESENT RECORDS AND DIGITAL ARCHIVES IN THE FRENCH ADMINISTRATION: OUR IT APPROACH

The team in charge of the development of this new system has particularly focused on archival description and metadata management because one of the major challenges is the description of digital archives and the capability to make requests in the new system over the time. One of the most important digital archives tested is mailbox and especially emails. Real examples of the Ministry of Defence and the Ministry of Culture were used to request the NoSQL database [4].

Digital archives lead to two difficulties: the number (and indirectly the raw volume) and the diversity. These two aspects are essential to ensure the performance and ease of interfacing of the solution [5].

Because of its volume and complexity, a platform for digital archiving leads to a significant cost. Two approaches have existed so far:

- **The vertical model**: each business archive theme is associated to one dedicated repository. This approach, functionally easy to implement, has the disadvantage of multiplying investment costs (one platform per business model). One main reason is the metadata format is different for each business, leading to a dedicated pattern of database model. Those solutions use the standard relational database model that provides performance, sometimes volume ability, but at the expense of absolute unification of the metadata representation model, leading to one dedicated silo per profession.
- **The horizontal model**: the platform is seen as a secure storage space but metadata are missing due to their diversity, hence their lack of control and completeness. These solutions use a simplified metadata format (mainly technical aspects) and therefore also rely on standard relational database model, deporting business record management to the relevant business applications. The advantage, compared to the vertical model, is the sharing of storage infrastructure and preservation method among various business domains. The disadvantage is the dissociation of storage and metadata, since the business metadata are kept on the business side.

In the VITAM project, we identified one core shared development part as the "back-office" managing all properties of an EAS (Electronic Archival System), not related to business or organizational aspects ("front-office" applications). However, in order to be effective and sustainable, this core must carry the indexing and search functions related to metadata, either on business, technical or archivistic fields. The main reason is metadata are just as important as the records themselves and must therefore be stored in the EAS too.

In addition, we have seen that the "front-office" need is not to store the metadata, but to know their structure (business model) and to have the ability to query them. Thus we propose to fully integrate the metadata query feature in the EAS, but to leave the control of the requested data (model) to the front-office. This leads to large data model variety capability.

In the context of "big" archiving system (several billion entries), another problem occurs, related to the ability to manage a huge database containing the metadata (several TB or even hundreds of

TB), while maintaining good performance, to ensure proper platform sharing, and of course the ability to grow as needed.

To meet these twin problems, but also to a single (either big volumes, either multiple data models), we proposed to study the use of document-store NoSQL database model [6] which has the main following properties:

- Ability to handle high volumes;
- Ability to handle flexible patterns (a table can contain multiple representations of the data in JSON-like format);
- Ability to provide high availability;
- Ability to provide high performances;
- Ability to handle custom queries;
- Ability to deal with full text requests.

3. A NEW DATA MODEL TO DESCRIBE ARCHIVES

The representation model is schematically presented as follows, inspired from MoReq 2010 [7].

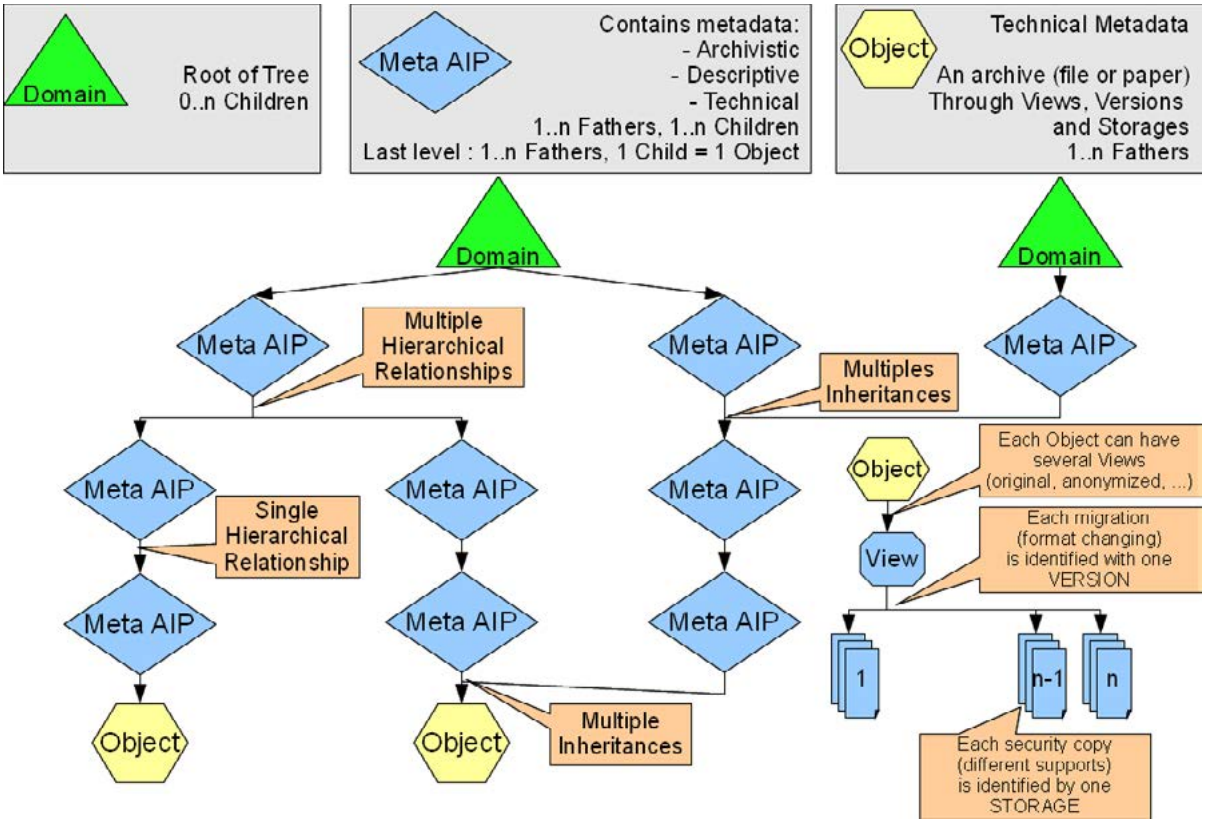


Figure 1. VITAM Data model inspired by MoReq2010. France, Prime Ministry, DISIC/POC, part 1, April 2013.

It is important to underline that this is an implementation experiment of the interesting data model of MoReq2010 in the French Archives administration (for records and digital archives).

In addition to the standard model, the ability to have multiple inheritance for each node in the graph, while not allowing cycle, leads to a directed acyclic graph (DAG as defined in mathematical theory, for instance in [8]).

The impact of multiple inheritances (multiple parents) is that inherited properties can have multiple values, due to multiple parents. While processing the search, the property resolution follows the access path to an object, since access to an object (or a meta-AIP or AIP or a node in the graph) is always from a root and following a path down to it.

The DAG approach is already used for medical ontology [9] and with RDF (Semantic Web) [10], but its application to archive classification scheme is quite new.

The structure of our DAG is as follow:

- A *Domain* is the root of a tree. There may be multiple roots, and one node may be accessed from multiple roots.
- A *Meta-AIP* is a node in the tree corresponding to a level in the classification scheme. It must contain enough information to be a good candidate.
- An *Object* is a node denoting an archive object. It is the smallest unit in a classification scheme (*Item*). It contains mainly technical information. In the case of a joint solution for paper and electronic, it is the lowest node for a paper archive, containing the location and packaging information.
- A *View* is a node to distinguish between different types of object representation from:
  - *Original* archive: the authentic piece, according to the original;
  - *Anonymized* view: similar to the original but with all the data relevant to privacy protection legislation withdrawn (i.e. ready for broadcasting);
  - *Raw* view in plain text format, useful for full text search or mixed presentation mode, for picture (scanned papers) and plain text formats for instance.
- A *Version* is a node to distinguish different versions of a View, following file preservation process (file format changing over time).
- A *Storage* is a copy of a version. It contains information about physical access to the actual archive. This is the lowest node in the hierarchy.

4. THE PROOF OF CONCEPT FOR THE IT DIRECTION OF THE PRIME MINISTER: NOSQL TECHNOLOGIES FOR DIGITAL ARCHIVING PLATFORM, A NEW APPROACH

To ensure the adequacy of this approach, we achieved a proof of concept based on an experiment from medium to large scale (a few hundred of GB to tens of TB) for metadata only. This article presents a subset of the results.

The objectives of this NoSQL study applied to archive metadata are as follows:

1. Ensuring that data model for representing metadata records is feasible and queryable;
2. Ensuring usage of flexible patterns is effective and practice;
3. Ensuring the performances in writing, but especially in reading are valid (ingest, access and preservation functions);
4. Ensuring these performances are met for a multiple concurrent clients ("front-office");
5. Ensuring the high availability of the solution and its robustness.

Firstly, the IT department of the Ministry of Culture made a study of NoSQL databases. Then, the experiment was done with real XML format of digital archive metadata on virtual machines (VM) at the IT Centre of the Ministry of Culture.

Each VM has 2 vCPU, 16 GB of RAM and 1 TB of disk. Up to 8 VM (x2 for reliability test) were created. The used softwares were MongoDB (version 2.4.3 http://www.mongodb.org/) for the NoSQL document database and ElasticSearch (version 0.90 http://www.elasticsearch.org/) for the indexation engine.

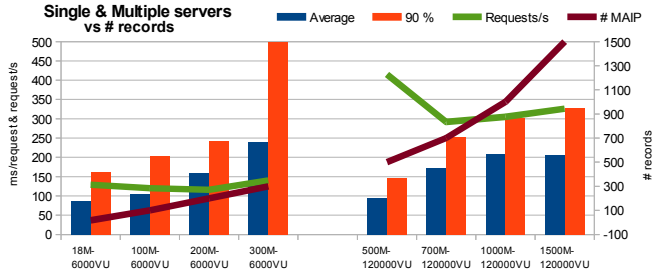


Figure 2. Single and multiple servers requests vs nb of records

The highest ingest performance was 7000 items/s with 8 VM, which leads to great DRP capability (less than 3 days for 1.5 billions of items), while this result is 4 time faster than with a single VM.

The graph 2 shows the metadata request performances on a single server (left side, 6 000 simulated users) and on multiple servers (right side, 120 000 simulated users), with up to 1.5 billion items with less than 500ms per request (90% of all times).

This graph shows that performances are still correct up to 300 millions on a single VM, even if the memory limit is reached. It shows also the good horizontal scalability, as previously observed during ingest (insert operations x4) but during access (request operations). We were able to grow 5 times bigger than with a single VM with 8 time servers.

During the tests, we used concurrently 10 different data structures (JSON schema) put in the same DAG without any issue, thanks to the schema-less capability of the NoSQL databases.

Finally, our reliability tests were also conclusive with no service interruption while disasters were simulated.

To conclude, the use of NoSQL technologies to cope with our needs of irregular description and variety of digital archives appears to be a perfect choice in term of performance, requests capabilities and adaptation to the digital administration and to the future digital information governance.

5. ACKNOWLEDGMENTS

Our thanks to Jean-Séverin LAIR, IT director of Ministry of Culture, for allowing us to build and to run our proof of concept in real conditions.

6. REFERENCES

[1] ISO 14-721:2003, OAIS, *Reference model for an Open Archival Information System*.

[2] ISO 16 175 (ICA-Req), CONSEIL INTERNATIONAL DES ARCHIVES, *Principes et exigences fonctionnelles pour l'archivage dans un environnement électronique*, Paris, 2008.

[3] FUENTES HASHIMOTO (Lourdes), Projet VITAM, Dossier de conception générale, Partie 2, *Modèle fonctionnel et technique*, v. 1.2 du 25 juin 2013, p. 11.

[4] LAPERDRIX (Marie), VASSEUR (Edouard), Projet VITAM. *L'archivage des messageries. Preuve de concept VITAM/Volet 2*, v. 1.0 du 27 juin 2013, p. 71-72.

[5] For instance, the National Archives of France manage 300 millions of items in their digital platform today for 20 Tb. Every item is described (business and technical metadata) and could be accessible in the reading rooms of the National Archives of France. Every item is specific and need special technical and description treatments and preservation planning. CONCHON (Michèle), "Les 10 ans du système CONSTANCE", in *Gazette des Archives*, n°163, 4th trimester 1993.

[6] LITH (Adam), JAKOB (Mattson), *"Investigating storage solutions for large data: A comparison of well performing and scalable data storage solutions for real time extraction and batch insertion of data"*, Göteborg: Department of Computer Science and Engineering, Chalmers University of Technology, 2010, p. 70. "Carlo Strozzi first used the term NoSQL in 1998 as a name for his open source relational database that did not offer a SQL interface[...]"

[7] DLM Forum foundation, European standard MoReq, *Model Requirements for the Management of Electronic Records,* 2011, available online : http://moreq2010.eu/pdf/moreq2010_vol1_v1_1_en.pdf

[8] CHRISTOFIDES (Nicos), *Graph theory: an algorithmic approach*, Academic Press, 1975, p. 170–174.

[9] SUPERKAR K. et al., Knowledge zone: *A Public Repository of Peer-Reviewed Biomedical Ontologies*, *Medinfo 2007: Proceedings of the 12th World Congress on Health (Medical) Informatics,* Klaus A. Kuhn, James R.. Warren, Tze-Yun Leong, [Brisbane, Australia, 20-24 August 2007], p. 813.

[10] AKIYOSHI Matonoy, TOSHIYUKI Amagasay, MASATOSHI Yoshikawaz, SHUNSUKE Uemuray, *A path-based relational RDF database*, *Proceeding ADC '05* Proceedings of the 16th Australasian database conference - Australian Computer Society, Inc. Darlinghurst, Australia Volume 39, 2005, p. 95-103.

Multimedia Collections Management

Cláudio Manoel Duarte de Souza
Bahia's Reconcavo Federal University - UFRB
Study Group and Laboratory Practice and Free Software and Multimedia - LinkLivre
Maestro Irineu Sacramento Street, Centro,
Cachoeira, Bahia, Brazil,
Post Code - 44300-000
(+55 75) 8811-1087
claudiomanoelufrb@gmail.com

Rubens Ramos Ferreira
Bahia's Reconcavo Federal University - UFRB
Study Group and Laboratory Practice and Free Software and Multimedia - LinkLivre
Coronel Garcia Street, Centro, Cachoeira, Bahia,
Brazil,
Post Code - 44300-000
(+55 75) 9172-6444
rubens.museu@gmail.com

ABSTRACT
Multimedia Collection Management Research Project developed by the Study Group and Laboratory Practice and Free Software and Multimedia - LinkLivre, linked to Bahia's Reconcavo Federal University/UFRB, quests both to identify and classify electronic components and hardware; software technical particularities, application code, analogical and digital file formats and media chemical compositions, such as DVDs, CDs, magnetic tapes, etc., whose structure works support for two multimedia's works storage and reproduction, one by the artist Fernando Rabelo and another one by Jarbas Jacome, both professors of Visual Arts at Bahia's Reconcavo Federal University. Based upon such an identification, possible practices might be pointed out for the establishment of multimedia collection preservation policy within a medium term. Finally, the research ending product will be accomplished by a database development, based on technical characteristics of the identified multimedia works.

Categories and Subject Descriptors
H.2.3 [**Database Management**]: Documentation.

General Terms
Documentation.

Keywords
Preservation, Digital Archives, Multimedia Collections.

1. INTRODUCTION
The present proposal is part of the identification of technical and theoretical downgrading regarding to practices applied to documentation and preservation of works produced and stored on media and digital media1, safeguarded in Brazilian museum spaces.

According to data collected by the Brazilian Institute of Museums - IBRAM (2011) [1], Bahia has mapped about 152 museums, among which 72.9% have collections of visual arts. However, among the identified museums, only 48.1% of these institutions have control over the documentation of their collection, with only 25.9% of museums in Bahia has access to software cataloging management multimedia collections.

Upon identifying this lack of computerization and management of multimedia estates, the project of Multimedia Collection Management proposes the creation and the availability of a free use database, subjected to adjustments to institutional realities in Bahia State's museum spaces. Such a management tool is presented as a product of this research project which quests to collect data on main analogical and digital media preservation practices, liable to be used for multimedia estate storage and reproduction environment and support.

2. OBJECT OF STUDY
The multimedia artist and researcher Jarbas Jácome is master in Computer Science by Cin-UFPE. His works are produced in open codes (application), allowing other artists to (re)use those codes, generating a network of collaborative Artmedia works, expanding the scope of his work - a sociocultural practice accessibility. The research involves the identification and classification of the installation known as Twilight of the Idols.

The researcher and visual artist Fernando Rabelo is master in Art and Imaging Technology at the School of Fine Arts of Minas Gerais Federal University. He has created works for internet as the interactive animation and Hiperface Insomnia_01 site. He was featured in 2005 FILE with the QWERTY Contact installation. And in 2006 he was invited to perform a residency at Medialab, Madrid where he attended the Interactive 06 with the De:echo project, within partnership with Rafael Marchetti. He participated of the Vrije Academie Household Programs and World Wide Visual Factory - in Deen Haag, Amsterdam in 2008 and 2009 in which he developed a system for panoramic projections and artistic and interactive applications such as live-performace "Flying Saucers" exhibited during the 5 Days-off Festival - Amsterdam.

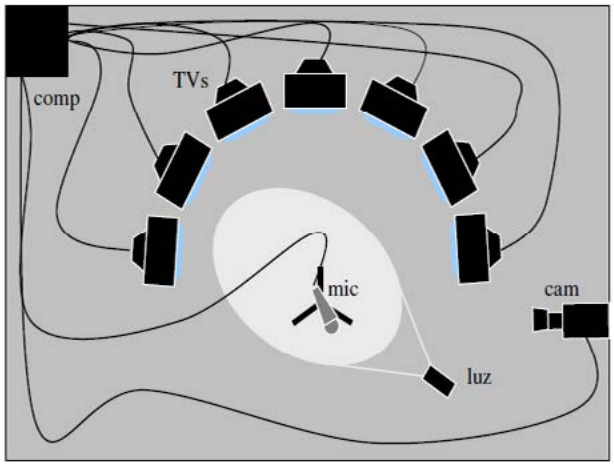
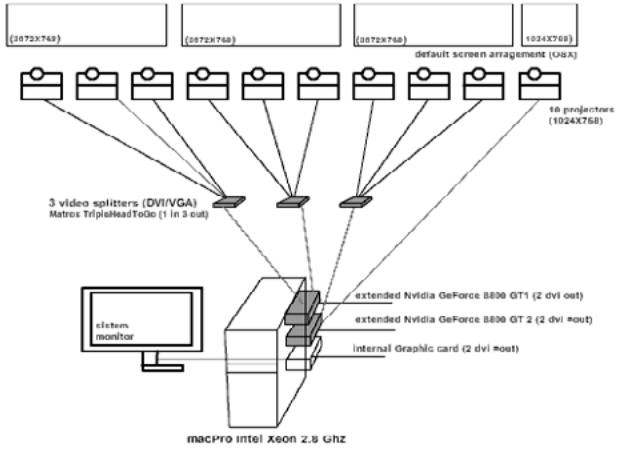


Figure 1. Top view of the installation Twilight of the Idols by Jarbas Jacome.

Both Rabelo and Jarbas use multimedia as a tool for audio, video and binary computing code dissemination in interactive networks. The processes turn to the immateriality of the relationships established in this space of virtuality. The explored interrelationships, identified in recent years in the languages of contemporary art, as in video art, conceptual art, happenings, installations, ready-made and socially engaged arts performances. Artistical languages that present new challenges to museum spaces related to cultural mediation - educational actions - as well as curatorship and estates preservation policies.



2. Panoramic Research Installation Structure.

3. RESEARCH BACKGROUND

Substantial portions of international museum institutions have departments and specific methodologies due to the preservation of multimedia works, such as Guggenheim Museum, London's Tate Gallery and San Francisco's Museum of Modern Art (Jill Sterret). Such spaces have qualified researchers staff who go in for more appropriate alternative studies applied to this estate typology preservation. Among the developed activities in art work field preservation that integrate digital formats in their production and reproduction, we have emphasized The Variable Media Network.

The program coordinated by Jon Ippolito and Alain Depocas, the Centre for Research and Documentation (CR + D) Daniel Langllis Foundation Director, aims to present recommendations for the documentation and art preservation, whose shapes tend to one almost instant obsolescence of the used technology. The mentioned recommendations are gathered in the Variable Media Questionnaire (VMQ), also planned by John Ippolito. The VMQ is an informational tool, similar to an interactive quiz where researchers, artists and museum professionals present problems and possible solutions concerning the preservation of multimedia works.

In Brazil, although it is possible to point out major advances related to the practices of art work preservation, formats and digital media which compose Brazilian collections are very poorly known. Similarly, we have realized how insignificant is the incentive to the discussion of strategies for the preservation of works in multimedia language. Until the present time, the Brazilian Institute of Museums - IBRAM, the highest t organ of representation for the museum sector in Brazil, has not yet presented a either a short or a medium term plan, not even within publications aimed at meeting the growing demand for managing multimedia collections taking into consideration our museological hybrid realities. It has become imperative and even emergencial, the incentive and immediate fostering to this reflection, as well as the proposal concerning multimedia language works preservation strategies, considering the changes that our media files have been subjected to. The absence of these incentives, technical and conceptual, might result in irreversible losses.

4. PROJECT METHODOLOGY

The Project Strategic Planning takes into consideration three basic conductive themes, as described below:

THEME I - Technical support applied to multimedia: It is understood as activity related to this theme, the classification of the specific techniques of analogical and digital media, such as computers, hardware, digital formats and application codes, Interactive Panoramic multimedia productions by Fernando Rabelo and Twilights of the Idols by Jabas Jacome.

Such data will be collected from the documentation provided by the artists themselves (work design project,) and bibliographic queries.

THEME II – Knowledge of the multimedia chemical compositions: Theme which comprises the identification of chemical agents, degradation and digital formats present in the storage and reproduction of the two selected works.

THEME III - Information management, database rearing and feeding: Theme that includes the development of a tool for managing multimedia collections (Database), based on the features identified in the classification of the two selected works. This tool comprises three fields of metadata management: Data on administrative management - GADM: This information comprises data on legal and administrative issues related to multimedia work.

Data for arts management - GART: This information comprises data on conceptual issues related to the work.

Data for the technical management - GTEC: These data include information on the technical issues (technology and preservation of supports/means) related to the work.

This way, the project is characterized as an innovative and dynamic initiative, since it establishes a level of accessibility and dissemination of knowledge aimed at the field of museum before limited by logistical issues, technological or even cash - assuming the constitution and territorial sociocultural realities, educational and economic status of Bahia State's municipalities, our current field. Similarly, possible means to create the use of a database suitable for this type of estate.

However, the use of such a management tool is not limited only to museological institutions, considering it can be applied to several types of cultural spaces, such as galleries, memorials, exhibition spaces and other environments that present estates of multimedia works, besides serving as a support tool to additional museum professional training as well as current students and museum specialists.

5. REFERENCES

[1] Museums in Numbers. Brasilia: Brazilian Institute of Museums, 2011. DOI=http://www.museus.gov.br/publicacoes-e-documentos/museus-em-numeros/

Author Index

CONFERENCES

iPRES 2004 Beijing, China, July 14 - 16, 2004
iPRES 2005 Göttingen, Germany, September 15 - 16 , 2005
iPRES 2006 Ithaca, NY, U.S.A, October 8 - 10, 2006
iPRES 2007 Beijing, China, October 11 - 12, 2007
iPRES 2008 London, United Kingdom, September 29 - 30, 2008
iPRES 2009 San Francisco, CA, U.S.A., October 5 - 6, 2009
iPRES 2010 Vienna, Austria, September 19 - 24, 2010
iPRES 2011 Singapore, November 1 - 4, 2011
iPRES 2012 Toronto, Canada, October 1 - 5, 2012
iPRES 2013 Lisbon, Portugal, September 2 - 6, 2013

ISBN 978-972-565-493-4