iPRES 2010

Proceedings of the 7th International Conference on Preservation of Digital Objects

7th International Conference on Preservation of Digital Objects
September 19 – 24, 2010 • Vienna, Austria

books@ocg.at
BAND 262

Gedruckt mit Förderung des Bundesministeriums
für Unterricht, Kunst und Kultur in Wien.

iPRES 2010 is organised by

Vienna University of Technology

Austrian National Library

Austrian Computer Society

Wissenschaftliches Redaktionskomitee

o.Univ.Prof.Dr. Gerhard Chroust
Univ.Prof.Dr. Gabriele Kotsis
Univ.Prof. DDr. Gerald Quirchmayr
Univ.Doz.Dr. Veith Risak
Dr. Norbert Rozsenich
o.Univ.Prof.Dr. Peter Zinterhof
Univ.Prof. Dr. Jörg Zumbach

Andreas Rauber, Max Kaiser, Rebecca Guenther, Panos Constantopoulos (eds.)

iPRES 2010

Proceedings of the 7th International Conference on Preservation of Digital Objects

iPRES 2010 gratefully acknowledges support from the following sponsors:

Gold Sponsors:

Silver Sponsors:

Bronze Sponsors:

Supporters:

Komitee für Öffentlichkeitsarbeit
www.ocg.at

Druck: Börse Druck
www.boersedruck.at

ISBN 978-3-85403-262-5

Conference Committee

General Chairs:

Andreas Rauber (Vienna University of Technology, Austria)

Max Kaiser (Austrian National Library, Austria)

Programme Chairs:

Rebecca Guenther (Library of Congress)

Panos Constantopoulos (Athens University of Economics and Business, Greece; Digital Curation Unit, Greece)

Panel Chair

Heike Neuroth (Göttingen State and University Library, Germany)

Tutorial Chair

Shigeo Sugimoto (University of Tsukuba, Japan)

Workshop Chairs

Perry Willett (California Digital Library, US)

John Kunze (University of California, US)

Local Organisation Chair

Johann Stockinger (Austrian Computer Society, Austria)

Publicity Chairs

Priscilla Caplan (University of Florida, US)

Joy Davidson (University of Glasgow, Scotland)

Local Organising Committee

Christoph Becker (Vienna University of Technology, Austria)

Jakob Frank (Vienna University of Technology, Austria)

Ellen Geisriegler (Austrian National Library, Austria)

Michael Greifeneder (Vienna University of Technology, Austria)

Mark Guttenbrunner (Vienna University of Technology, Austria)

Michael Kraxner (Vienna University of Technology, Austria)

Hannes Kulovits (Vienna University of Technology, Austria)

Thomas Lidy (Vienna University of Technology, Austria)

Rudolf Mayer (Vienna University of Technology, Austria)
Michaela Mayr (Austrian National Library, Austria, Austria)
Edith Michaeler (Austrian National Library, Austria, Austria)
Petar Petrov (Vienna University of Technology, Austria)
Andreas Rauber (Vienna University of Technology, Austria)
Stephan Strodl (Vienna University of Technology, Austria)

Programme Committee

Reinhard Altenhöner (German National Library, DE)

Bjarne Andersen (State and University Library Arhus, DK)

Pam Armstrong (Library and Archives Canada, CA)

Andreas Aschenbrenner (Göttingen State and University Library, DE)

Christoph Becker (Vienna University of Technology, AT)

Jose Borbinha (Instituto Superior Tecnico, PT)

Karim Boughida (George Washington University, US)

Karin Bredenberg (Swedish National Archives, SE)

Adrian Brown (Parliamentary Archives, UK, UK)

Gerhard Budin (University of Vienna, AT)

Priscilla Caplan (Florida Center for Library Automation, US)

Gerard Clifton (National Library of Australia, AU)

Euan Cochrane (Archives New Zealand, NZ)

Paul Conway (University of Michigan, US)

Robin Dale (Lyriasis, US)

Angela Dappert (British Library, UK)

Joy Davidson (University of Glasgow, UK)

Michael Day (UKOLN, University of Bath, UK)

Janet Delve (University of Portsmouth, UK)

Raymond Van Diessen (IBM, NL)

Martin Doerr (Institute of Computer Science - FORTH, GR)

Jon Dunn (Indiana University, US)

Miguel Ferreira (University of Minho, PT)

Dimitris Gavrilis (Digital Curation Unit, Athena Research Centre, GR)

David Giaretta (Rutherford Appleton Laboratories, UK)

Andrea Goethals (Harvard University, US)

Emily Gore (Clemson University, US)

Rudolf Gschwind (Uni Basel, CH)

Mariella Guercio (Universita' degli Studi di Urbino Carlo Bo, IT)

Mark Guttenbrunner (Vienna University of Technology, AT)

Thomas G. Habing (University of Illinois at Urbana-Champaign, US)

Ann Hagersfors (Lulea Technical University, SE)

Matthias Hemmje (Fernuni Hagen, DE)

Jeffrey van der Hoeven (Koninklijke Bibliotheek, NL)

David Huemer (Secure Business Austria, AT)

Jane Hunter (University of Queensland, AU)

Angela Di Iorio (Fondazione Rinascimento Digitale, IT)

Greg Janée (University of California at Santa Barbara, US)

Leslie Johnston (Library of Congress, US)

William Kehoe (Cornell University, US)

Ross King (Austrian Institute of Technology, AT)

Pavel Krbec (Czech Technical University , CZ)

Hannes Kulovits (Vienna University of Technology, AT)

Brian Lavoie (OCLC, US)

Christopher A. Lee (University of North Carolina, US)

Bill Lefurgy (Library of Congress, US)

Jens Ludwig (Göttingen State and University Library, DE)

Maurizio Lunghi (Fondazione Rinascimento Digitale, IT)

Julien Masanes (european Web Archive, NL)

Nancy McGovern (ICPSR, US)

Andrew McHugh (HATII at University of Glasgow, UK)

Carlo Meghini (CNR- ISTI, IT)

Salvatore Mele (CERN, CH)

Ethan Miller (University of California at Santa Cruz, US)

David Minor (University of California at San Diego, US)

Reagan Moore (University of Chapel Hill, NC, US)

Jacob Nadal (University of California at Los Angeles, US)

Quyen Nguyen (US National Archives and Records Administration, US)

Achim Osswald (Cologne University of Applied Sciences, DE)

Evan Owens (Ithaka, US)

Christos Papatheodorou (Ionian University, GR)

Bill Parod (Northwestern University, US)

David Pearson (National Library of Australia, AU)

Mema Roussopoulou (University of Athens, GR)

Chris Rusbridge (Digital Curation Centre, UK)

Raivo Ruusalepp (Dutch National Archives, NL)

Lisa Schiff (California Digital Library, US)

Michael Seadle (Humbold University, DE)

Robert Sharpe (Tessela, UK)

Barbara Sierman (KB, NL)

Tobias Steinke (German National Library, DE)

Randy Stern (Harvard University, US)

Shigeo Sugimoto (University of Tsukuba, JP)

David Tarrant (Southampton University, UK)

Daniel Teruggi (Institut National de l'Audiovisuel, FR)

Manfred Thaller (University of Cologne, DE)

Susan Thomas (University of Oxford, UK)

Helen Tibbo (University of North Carolina, US)

Emma Tonkin (UKOLN, University of Bath, UK)

Hilde van Wijngaarden (KB, NL)

Richard Wright (BBC, UK)

Eld Zierau (Royal Library , DK)

Contents

Preface

We are happy to present the proceedings of the International Conference on the Preservation of Digital Objects (iPRES 2010). This is the 7th conference in this series, which was held in Vienna, Austria on Sept. 19-24 2010, following previous events held in Beijing in 2004 and 2007, Göttingen in 2005, New York in 2006, London in 2008 and San Francisco in 2009.

With the tremendous increase of activities in the community of researchers and practitioners in the field of Digital Preservation, iPRES2010 has grown from a 2-day event to a full week of activities including tutorials, panels, poster sessions, and post-conference workshops, addressing a vast range of topics.

The conference started off with a full day of tutorials addressing the following topics: The next-Generation JHOVE2 Framework and Application; An exploration of the PREMIS Dictionary for Preservation Metadata; Logical and bit-stream preservation integrated digital preservation using Plato and EPrints; Personal Digital Archiving; and Stability of digital resources on the Internet and strategies for persistent identifiers. This was followed by the main conference from Monday to Wednesday, featuring sessions on Metadata, Policies, Business Models, Preservation Planning, System Architectures and numerous Best Practice reports. Tony Hey from Microsoft Research opened the conference with a keynote address on the emergence of the so-called 4th Paradigm in scientific research, focussing on the importance of massive data collections and the role of research libraries in this field. Patricia Manson from the European Commission presented the second keynote, highlighting the need for cross-disciplinary approaches to tackle the challenges in digital preservation.

iPRES2010 furthermore featured two panel discussions addressing rather controversial topics. The first panel discussed different approaches to preserving data on the World Wide Web, highlighting the different institutional policies and their consequences on Web data preservation. The second panel took up the topic of green computing, analysing the potential of digital preservation.

Two additional plenary sessions provided first of all a means for all poster presenters to demonstrate their key message in a short poster-spotlight presentation, offering the audience a tour-de-force through numerous preservation activities all around the globe. A second plenary session adopted a concept introduced at iPRES2009: authors could sign up spontaneously for short "Lightening Talks", briefly presenting interesting research challenges to the community and discussing different approaches and best practices for given problems.

The main conference was followed by a set of 5 Workshops, including long-running events such as the 10th International Workshop on Web Archiving (IWAW2010); a workshop on Spanning the Boundaries of Digital Curation Education; the second PREMIS Implementation Fair; a workshop on Collaboration, cooperation and grand challenges in digital preservation; as well as the Expert User Group Forum for Heritrix Operators and Developers, highlighting the strong focus on Web Arhciving activities in this year's conference.

With iPRES2010, the conference moved a step further to establishing a solid Peer Reviewing process for conference contributions. For the first time, iPRES2010 launched a call for full paper submissions, which were then reviewed by 4 members of the Scientific Programme Committee. The main review process was followed by an on-line discussion period between the reviewers, which then led to the final acceptance or rejection decisions. In addition to the full paper submission, however, iPRES2010 also kept the tradition of allowing abstract-only submissions within a dedicated late-breaking results track which had a simplified review process, but still were evaluated by 3 to 4 members of the Programme

Committee. Due to the increase in activities, competition for presentation slots at iPRES2010 was tough in spite of the increased length of the main conference. Out of 75 submissions, only 47 were accepted for inclusion in the iPRES2010 programme (acceptance rate: 63%), out of which 13 were selected for poster presentation. Grouping by tracks, only 21 out of 39 submitted full papers were accepted (acceptance rate: 54% within the full paper track), of which 4 papers were short papers and 17 long papers (acceptance rate: 44% within the full paper track). The late-breaking results track, which was based on the evaluation of extended abstracts, contributed 15 papers to the overall programme.

An important part of all iPres conferences is the social programme, and iPRES2010 offered several events, facilitating intensive discussions in an enjoyable and informal atmosphere. This included a wine tasting of Austrian wines as part of the Ice Breaking Party on Sunday evening. A reception was hosted by the City of Vienna in the City Hall, and the Austrian National Library hosted the Conference Dinner, following a guided tour through the splendid State Hall.

While the excellent scientific programme of the conference was only possible due to the many great research and best practice papers submitted to the conference and the thorough selection procedure supported by the Scientific Programme Committee, this conference would not have been possible without the extensive support of our sponsors. We would thus like to thank specifically our Gold Sponsors Tessella and Siemens. We are also indebted to our Silver Level Sponsors ExLibris and JISC, as well as Bronze Level Sponsors NetApp and Microsoft Research. The Conference also received support from AustralianScience as well as the Open Planets Foundation. We thank all our sponsors and supporters for their contribution to this conference and to the Digital Preservation community. We would also like to thank all staff involved in organizing this conference, both as part of the international Organizing Committee, as well as specifically Eugen Mühlvenzl, Johann Stockinger, Elisabeth Waldbauer and Elisabeth Maier-Gabriel at the Austrian Computer Society; Ellen Geisriegler, Michaela Mayr, Heide Darling, Michael Kranewitter, Edith Michaeler, Michaela Rohrmüller and Sven Schlarb at the Austrian National Library, as well as Christoph Becker, Michael Greifenender, Mark Guttenbrunner, Michael Kraxner, Hannes Kulovits, Rudolf Mayer, Petar Petrov, Stephan Strodl, and Natascha Surnic at the Department of Software Technlogy and Interactive Systems of the Vienna University of Technology. They all invested tremendous efforts to make sure that iPRES2010 was an exciting, enjoyable and successful event. Although organizing iPRES2010 was hard work, it was also a pleasure to work with such a competent group of people.

Rebecca Guenther, Panos Constantopoulos
iPRES2010 Programme Chairs

Max Kaiser, Andreas Rauber
iPRES2010 General Chairs

KEYNOTE

THE FOURTH PARADIGM: DATA-INTENSIVE SCIENTIFIC DISCOVERY AND THE FUTURE ROLE OF RESEARCH LIBRARIES

Tony Hey
Vice President of External Research
Microsoft

ABSTRACT:

We see the emergence of a new, 'fourth paradigm' for scientific research involving the acquisition, management and analysis of vast quantities of scientific data. This 'data deluge' is already affecting many fields of science most notably fields like biology with the high through-put gene sequencing technologies; astronomy with new, large-scale, high-resolution sky surveys; particle physics with the startup of the Large Hadron Collider; environmental science with both new satellite surveys and new deployments of extensive sensor networks; and oceanography with the deployment of underwater oceanographic observatories. This revolution will not be confined to the physical sciences but will also transform large parts of the humanities and social sciences as more and more of their primary research data is now being born digital. This new paradigm of data-intensive scientific discovery will have profound implications for how researchers 'publish' their results and for scholarly communication in general. The details both of what will need to be preserved and how this will be accomplished to create an academically valid record of research for the future are only now beginning to emerge. What is clear, however, is that research libraries have the opportunity to play a leading role in this ongoing revolution in digital scholarship. Repositories for both text and data are certain to play an important role in this new world and specialists in semantics, curation and archiving will need to work with the different research communities to fulfill their needs. Relevant projects and key collaborations recently undertaken by Microsoft Research will be highlighted, as will other Microsoft efforts related to interoperability and digital preservation.

Other resources:

- Link to Tony's photo(s) and short-bio:
http://www.microsoft.com/presspass/exec/tonyhey/default.aspx
- Link to Tony's full CV and links to talks:
http://research.microsoft.com/en-us/people/tonyhey/

KEYNOTE

DIGITAL PRESERCATION RESEARCH: AN INVOLVING LANDSCAPE

Patricia Manson
European Commission
Acting Director
Digital Content and Cognitive Systems

ABSTRACT:

One irony of the information age is that keeping information has become more complex than it was in the past. We not only have to save physical media and electronic files; we also need to make sure that they remain compatible with the hardware and software of the future. Moreover as the volumes of information, the diversity of formats and the types of digital object increase, digital preservation becomes a more pervasive issue and one which cannot be handled by the current approaches which rely heavily on human intervention. Research is needed on making the systems more intelligent.
For the research community, the challenge is also to build new cross-disciplinary teams that integrate computer science with library, archival science and businesses. We need to ensure that future technology solutions for preservation are well founded and grounded in understanding what knowledge from the past and from today we need to keep for the future.

Session 1a: Metadata and Object Properties

FULL PAPER

DEAL WITH CONFLICT, CAPTURE THE RELATIONSHIP: THE CASE OF DIGITAL OBJECT PROPERTIES

Angela Dappert
The British Library
Boston Spa, Wetherby, West Yorkshire LS23 7BQ
UK

ABSTRACT

Properties of digital objects play a central role in digital preservation. All key preservation services are linked via a common understanding of the properties which describe the digital objects in a repository's care. Unfortunately, different services deal with properties on sometimes different levels of description. While, for example, a preservation characterization service may extract the *fontSize* of a string, the preservation planning service may require the preservation of the text's *format*ting. Additionally, a value for the same property may be obtained in various ways, sometimes resulting in different observed values. Furthermore, properties are not always equally applicable across different file formats.

This report investigates where in these three situations relationships between properties need to be defined to overcome possible misalignments.

The analysis was based on observations gained during a case study of the nature of the properties that are captured in different institutions' preservation requirements and those of use in Planets preservation services.

1. INTRODUCTION

Planets [7] is a four-year project co-funded by the European Union to address core digital preservation challenges. In the Planets project, we have been developing a tool set of digital preservation services. Properties of digital objects play a central role in how these digital preservation services co-operate. All key preservation services are linked via a common understanding of the properties which can be used to capture the description of a digital object in a repository's care [5]. Unfortunately, we observe that different services tend to express the properties at different levels. There is, for example, a gap between the properties extracted by typical tools and the properties that stakeholders use to express their preservation requirements. It also has been observed that values for properties may be obtained in different ways; this may result in different observed values. Additionally, inherent differences between file formats make the comparison of some properties difficult.

In this paper, we analyse preservation plans and preservation services to determine what sorts of properties are expressed. We categorize how their values can be obtained. Each category determines property values in a particular functional or relational way. We illustrate the categories with real-life examples.

This work impacts practitioners, researchers and tool developers. The analysis shows where we can push the boundaries of automation to compute properties. It supports the argument that incomplete, approximate and heuristic values need to be accommodated. It illustrates why there is a need for an expression language for properties to define derived properties. It also illustrates why there is a need for robust aggregate comparisons of digital object property values. Finally, it argues that there is a need to capture the semantics of similar properties.

1.1. Preservation Services that Use Digital Object Properties

Preservation services that use digital properties (see Figure 1) include

- characterization services, such as the XCL services [16] or JHOVE [1], use file format knowledge to extract property values from digital objects in order to describe them. They may, for example, determine the dimensions of an image file.
- Testbed services, such as the Planets Testbed service [2], derive statistics on the performance of preservation action services, such as those performed by a file format migration tool. They determine to what degree those services preserve properties for representative corpora of digital objects. They, for example, measure the degree to which a service preserves *imageWidth* by evaluating it on many object migrations.
- Preservation monitoring services of the future will determine when a preservation risk for a digital object has arisen and trigger preservation planning.

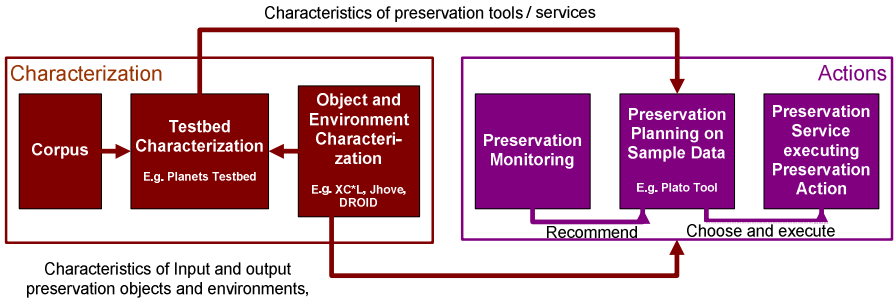


Figure 1: Digital Preservation Services

- Preservation planning services, such as Plato [3], determine which preservation action workflow best preserves the significant characteristics[1] [6] of a sample object set and issue a recommendation of action.
- Preservation action services, such as ImageMagick [10], execute migrations and other preservation actions on specific preservation objects and environments.

1.2. Approaches to Describing Digital Object Properties

In order for these services to work together, they need a common definition of properties. This is necessary in order to refer to properties unambiguously and to ensure interoperability and exchange across not only services, but also systems and institutions. In the preservation community, the definition of digital object properties is currently supported through the following approaches.

- Registries, such as Pronom [17], record properties that are applicable to a given file format, together with data constraints or a controlled vocabulary.
- Preservation metadata dictionaries, such as PREMIS [13], define common preservation metadata elements to describe properties of digital objects or their environments, together with data constraints or a controlled vocabulary, in a file format independent way.
- The InSPECT project [11] identified properties that apply to content types, such as images or emails, rather than to file formats.
- Controlled vocabulary registries, such as the Authorities and Vocabularies service of the Library of Congress [12], capture these properties' permissible values.
- Since related properties are often not immediately comparable, it is useful to develop a properties ontology which captures not only properties of digital objects but also describes them and the relationships between properties explicitly. The Planets Property Ontology is an example. A subset of it, the XCL ontology, is described in [14]. The issues discussed in this report illustrate why such a rich description of digital object properties is needed.

[1] In this paper "property" refers to an abstract trait of a digital object, while "characteristic" refers to a property / value pair of a concrete digital object.

2. POSSIBLE PROPERTY CLASHES ACROSS SERVICES

Different preservation services deal with properties on different levels of description. While, for example, a preservation characterization service may extract the *fontSize* of a string, the preservation planning service may require the preservation of the text's *formatting* in general. These properties may be related in interesting ways and are not comparable through simple equalities.

As a first generation proposal, Heydegger [8] outlines a framework of how property differences between preservation planning and preservation characterization services might be reconciled. This problem deserves generalized development resulting in both theoretical and practical solutions.

2.1. Preservation Services Interactions

Clashes between preservation services may show up in the following situations.

2.1.1. Preservation Planning and Preservation Actions

Stakeholders specify significant characteristics [6] of their preservation objects that need to be preserved (or obtained) through a preservation action. Preservation planning and preservation action services need to determine reliably whether these significant characteristics have been preserved. They request the values for the properties mentioned in the significant characteristics from the preservation characterization service. The characterization service is supposed to deliver the values for these properties in the required way. The preservation planning service additionally requests characteristics that describe the preservation action tools' performance from the testbed service in order to select tools that suit the sample data. These also need to align with the properties expressed in the significant characteristics.

2.1.2. Preservation Monitoring

Policy documents can specify which characteristics of digital objects and their environments manifest a preservation risk. In order to determine whether an object is at risk the monitoring service requests the object's characteristics from the characterization service. The properties used by the two services need to align.

2.1.3. Testbed Experimentation

During a testbed experiment, a preservation action service is tested on a set of digital objects, called a corpus. During the test, derivative objects are created whose property values are compared to the property values of the original objects. The results of this comparison describe the behaviour of a preservation action service based on the degree to which the service preserves the properties' values. There are two possible clashes. Firstly, this result is only meaningful if the testbed tests for a set of properties that are relevant to the users, whose requirements are captured by preservation planning services. Therefore the properties used in preservation planning and those tested in the testbed should align. Secondly, the testbed needs to obtain values of the measured property from preservation characterization services and their properties need to align.

Additionally, the testbed needs to aggregate test results that describe tool characteristics (rather than object characteristics) in a way that is most meaningful to their users and write them to a registry ready for use. Preservation planning services weigh those service characteristics to determine the optimal service for the users' specific preservation needs. The properties used by both need to align.

2.1.4. Corpus Design

A corpus is a set of digital objects with known characteristics for use in experiments. In order to compile benchmark corpora on which one can run testbed experiments in a representative way, one has to have an understanding of the applicable and relevant properties. Testbed results are meaningful to preservation planning services only if they are derived on a corpus of digital objects that reflects real life applications and contains instances of all properties that are relevant to users. It is, therefore, important that a corpus covers all properties that might be expressed by users in significant characteristics.

2.1.5. Preservation Action Tool Enhancement

Developers of a migration tool would like to ensure that a digital object after migration with this tool has the same properties as the digital object before migration. To achieve this they specify which property of the source format is to be transformed into which property of the target format. They then migrate sample files and test whether their assessment of property relationships was accurate and whether the migration tool maintained the properties faithfully. The properties of the source and target file format need to align.

They may also ask human subjects to assess the degree of conformance of the target to the source object. The properties that the human subjects apply are not necessarily the properties which where defined by the tool developers. In this case corrections of the property relationships and of the tool are necessary.

2.2. Stakeholders of Digital Object Properties for Preservation Purposes

The stakeholders interested in digital object properties are

- Creators of file formats who need to know how to design file formats so that properties of file formats can be reliably and consistently implemented across supporting applications, can be easily extracted, and validated, and can be migrated to different file format representations without damaging the content.
- Creators and curators of files who need to know which file formats have reliably determinable characteristics.
- Users of files who need to know how well validated a file is after undergoing a preservation action.
- Preservation policy officers and preservation plan developers who need to know which significant characteristics should be specified in their policy documents and validated reliably.
- Migration tool developers who need to know which characteristics to use in order to measure the authenticity delivered by their migration tool.
- Characterization tool developers who need to know how to extract characteristics or infer them from others.
- Testbed, corpora, preservation action and planning services developers, who need to know which properties can be obtained and which are required by users.

3. POSSIBLE PROPERTY CLASHES ACROSS VALUE ORIGINS

During research within the Planets project we observed that the values of digital object properties can be obtained in several ways. This section suggests an initial categorization of their value origin. It shows

- how the value for the same property can be obtained in different ways, possibly resulting in clashing, observed values.
- how different properties can be related to derive one property's value from others. This can help to mitigate the property clashes described in the previous section.

3.1. Value Origins

3.1.1. Extractable, File-Based Value Origins

Category description:
The value origin is a function of the simple digital object: f(object).

The original source of values may be a file, byte-stream or bit-stream. Values are extracted using a tool which implements an algorithm. For effective, scalable preservation, the tool would support *automatic* extraction of properties.

Examples:

- *imageWidth*
- *colourSpace* in PNG and other formats
- *linkURLs* in HTML
- *numberOfAudioChannels*

Derivability:
Algorithms for value extraction are based on file format specifications. This category is implemented for basic file-format-based properties in preservation characterization services, such as the XCL services [16] or JHOVE [1].

3.1.2. Extractable, Complex Value Origins

Category description:
The value origin is a function of a complex digital object and/or the object's environment:
f (object$_1$, …, object$_n$, environment).
These are property values that cannot be taken from the file alone, but rather need to be extracted from

- a representation – that is, the set of files that makes up one complete rendition or execution of a digital object (such as an HTML file with its embedded JPG files).
- a representation including auxiliary files (such as style sheets, non-embedded fonts, java scripts in HTML files, and schema definitions).
- the whole rendering stack (i.e. the preservation object's processing and presentation software and hardware environment).

These properties are not captured in a file format specification alone but are based on the whole environment as depicted in Figure 2.

Examples:

- A Microsoft Word document contains a link to a JPG file. One needs to look at both files to infer characteristics about the image's appearance in the document.
- The *colour* of a hyperlink in an HTML file is determined by the accompanying stylesheet. Both files need to be considered to characterize the *colour* of the hyperlinks.
- The presentation of an HTML file depends on browser settings or the choice of browser. Characteristics will vary depending on configuration.
- The actual layout of a Microsoft Word document on paper depends on the printer driver.
- *imageWidth* can be obtained from the rendering software, e.g. Adobe Photoshop.
- *fileSize*, since it depends on the operating system, is derived by asking the file system, rather than counting the actual bytes.

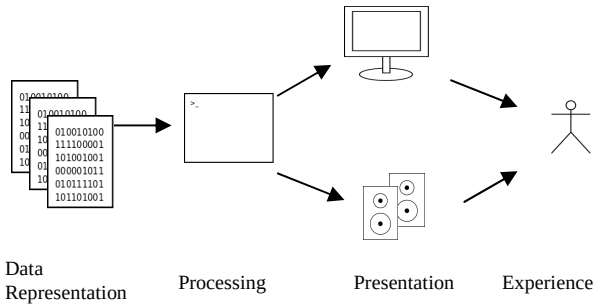


Figure 2. Digital Objects and Their Rendering Stack. (Adapted with permission from Jan Schnasse)

Derivability:
This is a generalization of characterizing one file at a time without regard to its environment. Once we include multiple files and environments into our scope, we expand the set of automatically extractable properties. This category could be implemented now. Some very useful information can be extracted easily; but some with, sometimes, considerable effort.

3.1.3. Non-Extractable, Complex Value Origins

Category description:
The value origin is a function that approximates the property's value
f' (complex object, environment) ≈ f (complex object, environment).

These are properties that are too complex to capture reliably in an algorithmic way, but they can be approximated by related metrics.

Examples:

- The stakeholders' observation of *imageQuality* does not always align with existing image quality metrics. But it is possible to define an acceptable metric which can be measured and compared [9].
- Different parameter configurations of *frequencies*, *amplitudes* and *modulations* can produce comparable sound to the human ear. Even if the representations are not identical, they can have an identical effect for the user. In this case, the property *perceivedSound* is an approximate metric which maps the measurable sound properties onto it.

- Pixel-wise different images may have the same effect on the human eye or rendering devices, since some differences cannot be perceived or rendered.

Multiple metrics can be created to define which combinations are perceived as the same *imageQuality*, *sound* or *colour*, respectively.

Derivability:
By definition, these characteristics cannot be inferred from extractable characteristics unless an algorithmically supported metric is developed. This category can be implemented now, but with, sometimes, considerable effort for development of the algorithmically supported metrics.

3.1.4. Implicit Semantics Value Origins

Category description:
The value origin is a heuristic that results in a value, as well as a confidence measure. The value and confidence measure are repeatable and always give the same results.
(f ' (complex object, environment, heuristic),
conf (complex object, environment, heuristic))

These are properties that require interpretation of semantics that is not captured in the preservation object and its environment. This can, for example, be achieved by employing knowledge-based heuristics.

Examples:

- Some CAD drawings of pipes only specify where pipes are, but not how they are connected. The connections may be clear to the user, but difficult to extract from the object and its environment.
- Older PDF formats do not have structural components such as titles, abstracts, footers. Even in newer PDF formats, functions supporting structural components are currently rarely used in practice during the document creation process. They can, therefore, not be reliably automatically identified.

Derivability:
Implicit semantics require knowledge-based reasoning to infer property values. The property values in this category can be determined reliably and repeatably, but with considerable effort.

3.1.5. Inferable Value Origins

Category description:
The value origin is a composite function of other value origins: $f(g_1 (object), \ldots, g_n (object))$.

These are properties that are not explicitly captured in the file format, but can be inferred from other properties. Values may be inherited in an object or property hierarchy, derived through a function from other values, or logically inferred.

This can also be used to relate properties that have synonymous names, by explicitly stating their equivalence.

Examples:

- *aspectRatio* of an image may be calculated as *imageWidth* / *imageHeight*
- *colourFidelity* can be measured from either of two different functions: *averageColour* or *histogramShape*
- *wordCount* can be measured in several ways: e.g., count hyphenated words as one or as multiple words
- *resolutionInPPI* can be mapped via its data type to *resolutionInLinesPerMillimeter*
- *imageWidth* of an image, used as property in one file format, may be inferred from the property *width*, used in another file format, by stating its equivalence with *width*.
- *bitDepth*, is described as one non-negative number in PNG and as three non-negative numbers (one per colour channel) in TIFF. Even though the property is the same in both cases, they have different data types for their values. This can in many cases be expressed through a functional relationship with which one can be derived from the other.

Derivability:
Algorithms for the value inference need to be defined. Even though this category can be implemented now, it has not widely been done. The property values in this category can be determined reliably and repeatably.
The specification of how the involved properties are related can be used to resolve clashes in levels of granularity between preservation services as discussed in Section 2.

3.1.6. Non-Predictable Value Origins

Category description:
The property value is always the same, but the observed value can be different at different times, for example due to interpretation.
f (complex object, environment, interpretation)

These are characteristics that possibly have different values when evaluated by different mechanisms (e.g. different people or the same person at different times).

Examples:

- *colourVibrance* can be judged differently by different observers.

Derivability:
The property values in this category can, by definition, not be reliably inferred.

For testbed purposes, the statistical average of these properties may well be determinable (See for example the Mean Observer Score metric [15].) But for the indi-

vidual digital object, these techniques can not be applied.

3.1.7. Time Varying Value Origins

Category description:
The property value is different at different times, depending on environmental changes. The observed value, therefore, can be different at different times.
f (complex object, environment, time)

These are properties whose characteristics cannot be reliably reproduced because of time varying behaviour / value change over time.

Examples:

- A time varying sequence of images in an HTML table cell, such as flashing advertisements, will result in different extracted images at different times.

Derivability:
The property values in this category can, by definition, not necessarily be repeatably inferred.

3.1.8. Indeterminable Value Origins

Category description:
The value can not be observed because the digital object is corrupted or the required knowledge is incomplete. In this situation, property values are not measurable at the time because you lack information.

Examples:

- An old Cyrillic font that is used in a document is not available on our machine configuration. An interesting discussion of this can be found in [4].

Derivability:
The property values in this category can, by definition, not be determined.

3.2. Property Categories that Are Independent of Digital Objects

There are additional property types that are independent of digital objects, but they still affect preservation services.

3.2.1. Representation Independent Properties

There are preservation properties that are independent of the file, representation or rendering stack.

There may, for example, be a requirement
"If a preservation action is chosen, it must be either a *migration* or a *data refresh*. Other preservation action types are not supported."

This requirement guides the preservation plan by specifying the property *preservationActionType*, but does not refer to properties which could be extracted from digital objects.

3.2.2. User Experience Properties

Different users experience (see Figure 2) the same performance of a digital object differently. E.g. somebody who participated in a competition will perceive images documenting the event different from somebody who was not involved or who does not understand the rules underlying the competition. Properties that describe the stakeholder's experience rather than the system's performance – those that relate to the psychological effect of object characteristics on a stakeholder - were not investigated within the Planets project.

This category is different from the *Non-Predictable Value Origins* category discussed in Section 3.1.6, since it considers emotional impact rather than how the value is obtained.

3.3. A Property Can Have Several Origins for a Value

If there are multiple ways of obtaining its value, a property can belong to several of the categories described in this section. E.g. *imageWidth* can be extracted from a file (category *Extractable, File-Based Value Origin*), calculated from other properties, such as *resolution* and *pixelCount* (category *Inferable Value Origin*), obtained from the rendering software (category *Extractable, Complex Value Origin*), or measured by hand from a printed sheet (category *Non-Predictable Value Origin*). *authorName* can be extracted from XML mark-up, HTML headers, MS Windows file properties, etc. (category *Extractable, File-Based Value Origin*) or entered by hand (category *Non-Predictable Value Origin*). *lineLength* can be extracted from a vector graphic (category *Extractable, File-Based Value Origin*) or calculated through heuristic algorithms based on a raster representation of the line (category *Implicit Semantics Value Origin*).

Whenever there are multiple origins for the value of a property there is a risk that there is a clash of the observed values and that they, therefore, represent a related rather than an identical property.

One important task of a property ontology is to capture those origins and their relationships.

3.4. Manually vs. Automatically Extracted Properties

Values for properties can be obtained automatically or manually. Much research has gone into automatically extractable properties. For large volumes of objects, manual declaration of property values by means of free format texts is unworkable. Unfortunately, it is evident that a large set of properties that users require can be extracted automatically only with great difficulty or not reliably. There is a justified desire, where possible, to capture relationships such that most characteristics can be automatically inferred from automatically extractable characteristics. However, as the *imageWidth* and au-

thorName examples illustrate, whether or not a property is obtained automatically is an orthogonal issue to our discussion.

3.5. Resolving Property Clash

Property ontologies have to deal with the semantics of similar properties so that they can be compared or derived from each other. This can be used to overcome the clashes between different preservation services that were observed in Section 2. From the preceding analysis, we observe that properties that are related to each other functionally (e.g. through a value origin definition in the *Inferrable Value Origins* category), can be related to each other through this definition within or across preservation services.

In all situations of clash, properties that are derived through non-repeatable value origins (e.g. through a value origin definition in *Non-Predictable* and *Time-Varying Value Origins* categories), cannot reliably be compared to other properties through simple equality metrics. They may be assessed with complex comparison metrics.

Properties that are non-determinable, e.g. in the *Indeterminable Value Origins* category, cannot be compared to others.

4. POSSIBLE PROPERTY CLASHES ACROSS FILE FORMATS

A key task of many preservation services is to compare properties of a digital object before and after a preservation action, such as a migration, in order to assess the quality of the preservation action. This may be hard to do due to incompatible file formats. This section discusses the reasons for this.

4.1. Properties for Different File Format Paradigms

4.1.1. Various Primary Components and Content Structures

Some related properties are hard to compare across file formats because those formats are represented in fundamentally different paradigms. Each file format has primary components. Properties apply to those components and are used to characterize a digital object of this file format. For example, a substring component of a text document can be described by the *fontType*, *fontColour*, and *fontSize* properties. When file format paradigms use different types of primary components, properties may not be easy to compare.

For example, both a Word document and a PDF document may represent the same text, but their underlying paradigms are quite different. PDF documents' primary components are representation elements, such as elements of the page layout. Their properties describe a fixed-layout 2D document with an underlying page orientation. Word documents' primary components are content elements, such as text strings, columns, or titles. Their properties describe them mostly independent of the page layout; for example, Microsoft Word has no notion of the page coordinate points where a paragraph starts. This results in a phenomenon where seemingly identical properties can actually refer to quite different properties. For example, the property *pageNumber* in Microsoft Word is determined by the author of the document. It may start with page numbering of a title page, or start after an introduction to the document. The PDF document displays page numbers starting with the first physical page. Even though it may display a different logical page number, it has no "awareness" of it.

Likewise, both vector graphics and raster graphics capture images. But while vector graphics describe the properties of content elements of the image (such as the *width*, *length* and *colour* of a line, or the *diameter* and *position* of circle), a raster image would represent the same content by recording properties of its representation elements, the pixels of the image. Raster image formats have no notion of properties of lines and angles; vector graphics formats have no notion of pixel properties.

Even though both the Open Document Format for Office Applications (ODF) and Office Open XML (OOXML) have content elements as primary components, their properties are not necessarily directly comparable because they use different models of how the text is structured. ODF uses a hierarchical content element decomposition into chapter, section, paragraph, marked up text, etc.. Properties apply to those structures. OOXML, however, applies its properties to runs of consistent mark-up which can span structural elements, for example, mark text as *bold* across paragraphs. In this case, one needs to not only capture the relationship between the properties, but also the relationship of the clashing structural elements.

Furthermore properties may cross content types, such as *image* or *text*. Font properties, for example, may cross text and image paradigms. Properties of fonts that are encoded as images cannot be easily compared to those of fonts that are encoded as characters.

4.1.2. Properties Describing Absolute and Relative Page Layout

In addition to differing primary components, file formats fundamentally differ by whether they have absolute vs. relative page layout. Of the example formats in this section, the image and PDF formats describe the absolute position of their content or representation elements, while Word and ODF documents describe the relative position of their content elements. Any properties describing positions on a page or positions of components relative to each other are hard to capture in their non-native representations.

4.1.3. Crossing File Format Paradigms

Which properties are easily extractable depends on the paradigm and primary components used. If one works within the paradigm of raster images, then pixel properties are easily extractable. From this perspective vector graphic elements are not easily extractable, and can, at best, be heuristically approximated. If one works within the paradigm of vector images, then graphic elements are the primary components with measurable properties. From this perspective, raster image pixel properties are not measurable.

Due to the inherent conceptual distance, shifting from one file format paradigm to another results in inaccuracies which make a reliable comparison based on properties hard. For example, one can convert a vector graphic into a raster image in order to compare it with another raster image to infer their similarities or differences. But the conversion algorithm does not necessarily produce a raster image that has pixel-wise equivalence to another raster image of the same content. This means that comparison metrics need to be developed that can anticipate the resulting inaccuracies while still capturing actual content differences.

4.2. Different Scope of Functionality of File Formats

Different file formats support different functionality. For example, OOXML has editing sessions, for which it records a modification and editing history. This functionality is not supported by some other file formats. It is therefore hard to compare properties relating to this differing functionality across file formats.

5. DISCUSSION AND CONCLUSION

This report investigates where in the preservation process interesting relationships between digital object properties occur that are not straight-forward to resolve. A property ontology is a way of modelling them explicitly in order to overcome possible misalignments.

The report suggests a categorization of how properties are obtained and discusses which of them can be used to resolve property clashes.

This work impacts practitioners, researchers and tool developers. The analysis shows where we can push the boundaries of automation to compute properties. It supports the argument that incomplete, approximate and heuristic values need to be accommodated. It illustrates why there is a need for an expression language for properties to define derived properties. It also illustrates why there is a need for robust aggregate comparisons of digital object property values. And it, finally, argues that there is a need to capture the semantics of similar properties.

From it we can develop a research roadmap into digital object properties for digital preservation tasks.

6. ACKNOWLEDGEMENTS

Work presented in this paper was carried out as part of the Planets project (IST-033789, http://www.planets-project.eu/) under the IST Programme of the European Sixth Framework Programme. The author is solely responsible for the content of this paper.

7. REFERENCES

[1] Abrams, S., Morrissey, S., Cramer, T. "“What? So What”: The Next-Generation JHOVE2 Architecture for Format-Aware Characterization." *The International Journal of Digital Curation*, Issue 3, Volume 4, 2009. http://www.ijdc.net/index.php/ijdc/article/viewFile/139/174

[2] Aitken, B. "The Planets Testbed: Science for Digital Preservation" *The Code4Lib Journal*, ISSN 1940-5758, Issue 3, June 2008. http://journal.code4lib.org/articles/83

[3] Becker, C., Kulovits, H., Guttenbrunner, M., Strodl, D., Rauber, A., Hofman, H. "Systematic planning for digital preservation: evaluating potential strategies and building preservation plans", *International Journal on Digital Libraries (IJDL)*, December 2009. http://www.ifs.tuwien.ac.at/ ~becker/pubs/becker-ijdl2009.pdf

[4] Brown, G., Woods, K. "Born Broken: Fonts and Information Loss in Legacy Digital Documents" *Proceedings of the 6th International Conference on Preservation of Digital Objects. iPres 2009*, 2009 http://www.cs.indiana.edu/~kamwoods/BrownWoodsiPRES09_Final.pdf

[5] Dappert, A., Farquhar, A. "Implementing Metadata that Guides Digital Preservation Services" *Proceedings of the 6th International Conference on Preservation of Digital Objects. iPres 2009*, 2009. http://www.planets-project.eu/docs/papers/Dappert_MetadataAndPreservationServices_iPres2009.pdf

[6] Dappert, A., Farquhar, A. "Significance is in the Eye of the Stakeholder" *European Conference on Digital Libraries (ECDL)*, September/October 2009, In: M. Agosti et al. (Eds.): ECDL 2009, LNCS 5714, pp. 297-308, 2009, © Springer-Verlag Berlin Heidelberg 2009 http://planets-project.eu/docs/papers/Dappert_Significant_Characteristics_ECDL2009.pdf

[7] Farquhar, A., and Hockx-Yu, H. "Planets: Integrated services for digital preservation" *Int. Journal of Digital Curation* 2, 2 (November 2007), 88–99 http://www.ijdc.net/index.php/ijdc/article/viewFile/45/31

[8] Heydegger, V., Becker, C. "Specification of basic metric and evaluation framework", Planets Project external deliverable PP5/D1. http://planetarium.hki.uni-koeln.de/planets_cms/sites/default/files/Planets_PP5-D1_SpecBasicMetric_Ext.pdf

[9] Heydegger, V. "Just One Bit in a Million: On the Effects of Data Corruption in Files" *European Conference on Digital Libraries (ECDL)*, September/October 2009, In: M. Agosti et al. (Eds.): ECDL 2009, LNCS 5714, pp. 315-326, 2009, © Springer-Verlag Berlin Heidelberg 2009

[10] Image Magick http://www.imagemagick.org/script/index.php

[11] Knight, G., Pennock, M. "Data Without Meaning: Establishing the Significant Properties of Digital Research" *The International Journal of Digital Curation*, Issue 1, Volume 4 | 2009 http://www.ijdc.net/index.php/ijdc/article/viewFile/110/87

[12] Library of Congress. "Authorities and Vocabularies". http://id.loc.gov/authorities/about.html

[13] PREMIS Editorial Committee "PREMIS Data Dictionary for Preservation Metadata, Version 2", March 2008. http://www.loc.gov/standards/premis/v2/premis-2-0.pdf

[14] Puhl, J. et alii "eXtensible Characterisation Language Suite" Chapter 4. Planets Report PC/2-D12; PC/2-D13; PC/4-D7. http://planetarium.hki.uni-koeln.de/planets_cms/sites/default/files/PC2D12D13PC4D7-01.pdf

[15] Reckwerdt, B. *Quantitative Picture Quality Assessment Tools*. http://www.videoclarity.com/WPUnderstandingJNDDMOSPSNR.html

[16] Thaller, M. *The eXtensible Characterisation Languages – XCL*. Verlag Dr. Kovač, Hamburg, 2009.

[17] The National Archives: PRONOM http://www.nationalarchives.gov.uk/pronom/

FULL PAPER

A METS BASED INFORMATION PACKAGE FOR LONG TERM ACCESSIBILITY OF WEB ARCHIVES

Markus Enders
The British Library
96 Euston Road
London, NW1 2DB

ABSTRACT

The British Library's web archive comprises several terabyte of harvested websites. Like other content streams this data should be ingested into the library's central preservation repository. The repository requires a standardized Submission- and Archival Information Package.

Harvested Websites are stored in Archival Information Packages (AIP). Each AIP is described by a METS file. Operational metadata for resource discovery as well as archival metadata are normalized and embedded in the METS descriptor using common metadata profiles such as PREMIS and MODS.

The British Library's METS profile for web archiving considers dissemination and preservation use cases ensuring the authenticity of data. The underlying complex content model disaggregates websites into web pages, associated objects and their actual digital manifestations. The additional abstract layer ensures accessibility over the long term and the ability to carry out preservation actions such as migrations. The library wide preservation policies and principles become applicable to web content as well.

1. INTRODUCTION

The web has become one of the primary information resources. Its information is read by the general public, cited by researchers and re-used by bloggers and commercial publishers. But information on the web is transient. Unlike printed books or journals information can easily be modified or deleted from electronic systems.

Since the mid 90ies when national libraries and the Internet Archive started archiving the web, the importance and the awareness for preserving the information published on web raised.

Today various web archives exists providing access to millions of web pages. The British Library's web archive contains more than 23516 instances of websites comprising of 5.5. TB of (compressed) data. As the size and use of the web archive grows, it becomes important integrating the web archive with the library's preservation system to ensure its long term availability. Therefore the library's preservation policies, supported preservation formats and preservation use cases must be considered.

2. WEB ARCHIVING

The British Library has set up a complex technical infrastructure for collecting, storing and providing access to web sites. A set of tools such as the Heritrix crawler, the Wayback Machine and the Web Curator Toolkit are used. These tools implement common interfaces and use the same file formats for exchanging data.

Legacy data in the archive had been harvested using the PANDAS system. It provides a different infrastructure for harvesting, storing and managing the web archive. As a result the file formats and interfaces are different.

From the long term preservation perspective it is not useful supporting two different formats for the same purpose. The SIPs and AIPs for the Digital Library System are standardized. The supported file formats are consolidated and unified.

2.1. Tools

2.1.1. PANDAS

As a member of the UK Web Archiving Consortium project, the British Library started very early to collect and store web pages. The only tool available at this time was the PANDAS system[1]. PANDAS provides all the required functionality for selecting, harvesting, managing and providing access to websites. As it was one of the first tools, it had scalability problems managing a huge number of concurrent crawls initiated by various curators. The output from PANDAS is very simple: Harvested bytestreams are stored in a directory

[1]http://pandora.nla.gov.au/pandas.html

structure in the local file system. The British Library did not capture comprehensive descriptive metadata nor logfiles.

2.1.2. Heritrix and the Web Curator Toolkit

The current solution for harvesting webpages consist of three components which provide an end-to-end process for harvesting, managing and disseminating archived webpages. It is much more scaleable than the format PANDAS solution.

The internet archive's Heritrix[2] crawler is used to harvest webpages. It starts with a one or more UDLS (so called seed-URLs), analyzes the received bytestream and extracts further URLs from HTML pages. Comprehensive configuration options as pattern matching for URLs, support of robots.txt and counting the crawl depth allows to restrict a single crawl to a specific area of the web. For the selective harvesting the crawl is usually restricted to a single web site.

The Web Curator Tool (WCT) is used by curators for managing the content and initiating crawls. The WCT allows the curator to set configuration options for the crawler and to schedule crawls for each target. A target is any portion of the web which the curator regards as important to collect and archive. Each target has at least one instance. This target instance is a snapshot of target at a particular point in time. Every time Heritrix is harvesting a target, a new instance is generated.

The Access Tool provides end user access to the harvested content. It uses the metadata which had been captured and generated by the Web Curator Tool as well as the content data which had been harvested by Heritrix. It integrates the open source version of the Wayback machine to access the individual bytestreams which had been harvested by Heritrix.

2.2. Data model of the Web Curator Toolkit

The OAIS model defines three different information packages for submission, archiving and dissemination. They provide the data which is needed to support the appropriate functionality. Information packages are an abstract concept which encompasses all the data being needed for a well defined set of functions. The information packages of current web archiving tools are focused on supporting submission and dissemination. The data structures and information are stored in a convenient way for the Access Tool and the Wayback Machine to disseminate the data.

Information packages can be split over various files, database records etc.

The web archiving toolset uses a so called ARC[3]container for storing the content data. It contains the actual bytestreams being returned as a result of every successful http-request. They are enriched with basic technical (size of the bytestream, mime type) and provenance (date and time of harvest) data. The ARC files are created by Heritrix. Every ARC file is accompanied by an index file. It allows non-sequential access, as it records the location of each URL within the ARC container.

Descriptive and rights metadata are not stored in the ARC container, but in a relational database. The information in the database is captured by curators using the Web Curator Tool.

The information in the The Heritrix crawler creates a number of different logfiles for each crawl. These logfiles contain provenance information about the harvested and stored bytestreams as well as those requests which failed. Failed requests may or may not return a bytestream. In case of http-errors the web servers are usually returns a bytestream and an appropriate error code in the http-header. This bytestream is stored in the ARC container. Other errors such as runtime errors of the software may not return a bytestream. The only evidence of such a request is recorded in the logfile. The logfile enables the curator to retrace the crawler's path through a website and discover the reason if the bytestream for some URLs is not available in the archive.

For provenance purposes the crawler's configuration is also stored. It stored the schedule for regular crawls which is set by the curators using the Web curator Toolkit. They are also responsible for configuring the crawler regarding the crawl depth and URL-patterns. These settings define the conditions for Heritrix to stop following hyperlinks. Besides this process related information, the Web Curator Tool allows the curator to capture descriptive metadata for each target. The metadata is used for resource discovery purposes.

According to the OAIS model both tools – PANDAS as well as WCT/Heritrix - are creating information packages which are used for submission (SIP) and dissemination (DIP). Their content model is defined by the technology being used and optimized for collecting and providing access to webpages. Long term preservation requirements had not been considered. Both SIPs support different standards and are structured differently. As a consequence the integration with other systems, including the library's long term preservation repository is very poor. The systems being used for web archiving are using their own technical infrastructure.

For preserving web content in the long term, the content model must be harmonized. The archival store can only support a single format for the Archival Information Package. This format must be based on common standards and consider long term preservation requirements. Besides the operational metadata being embedded in the SIP/DIP, additional archival metadata must be generated and stored.

[2] http://crawler.archive.org

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

3. PRESERVATION REQUIREMENTS

In the long term it will be difficult for the library to run and maintain different systems for storing, managing and preserving information. A library's Digital Library System (DLS) is responsible for storing Archival Information Packages from various sources and various content streams. As a consequence web content must to be ingested into the library's archival store as well. The same common standards and policies must be used.

The Digital Library System serializes metadata as well as content information as files in its internal file system. While the content is stored in so called container files, the descriptor of each package is serialized using the Metadata Encoding and Transmission Standard (METS)[4]. METS is a framework for describing a digital resource and all components of it. In this case it describes a single instance of a website harvested at a particular point in time (target instance). Every instance is stored in a separate information package.

The METS description of the resource comprises descriptive, technical and preservation metadata as well as the internal structure of the resource. The internal structure defines all the objects the resource consists of: abstract entities such as website and webpages, container files and bytestreams as well as their relationships.. METS uses so called extension schemas such as PREMIS, MODS, Dublin Core to store descriptive and preservation metadata Though a different schema is used, this metadata is part of the METS file.

The library's Digital Library Systems (DLS) stores the METS file and content files in its internal file store. The file store hold three distributed copies of every file. The actual content files are bundled in container files which are similar to the Heritrix output. Instead of ARC the standardized WARC format is used for the container files. All ARC containers had to be migrated to WARC prior to ingest as the DLS' ingest interface accepts only WARC containers with an appropriate METS descriptor.

Storing and preserving bytestreams is just one prerequisite for ensuring the accessibility of information in the future. File Formats, transport protocols and the supported software will change over the next decades. It is uncertain if web browsers and underlying HTML pages will still be the tools and formats of choice and how future software tools will be able to render today's web pages. Preservation actions such as emulation and migration will ensure that the information can still be rendered. Though the web page's manifestation (the bytestream) may change the appropriate documentation ensures information's authenticity.

Preservation metadata ensures the authenticity. It makes "digital objects self-documenting over time"[5] and is an important part of each Archival Information Package (AIP) record. It includes technical details on structure and format of a bytestream as well as the history of all actions taken to maintain the bytestream's information. It is part of the digital provenance metadata for each bytestream which is partly captured when it is harvested. Additional actions may occur prior to ingest into the repository: virus checks, format migrations or other transformations. These actions are properly documented in the preservation metadata record. In case preservation actions result in new bytestreams, the relationship between the old and new bytestream is recorded.

Though the AIP's data structure focuses on supporting preservation, it must consider dissemination use cases as well. Some functions of the access system rely on metadata which needs to be provided the AIP. According to the OAIS model, the Dissemination Information Package (DIP) is derived from the AIP.

3.1. Standardizing the SIP

The ingest process handles two different kind of submission packages. The PANDAS package is significantly different from the package provided by Heritrix and the WCT. Content data is stored in a directory structure instead of using container files. The metadata provided by PANDAS is very limited compared to the metadata the WCT provides.

For reasons of efficiency the British Library decided to standardize the submission information package for the purpose of ingest into the Digital Library System (DLS). Standardizing the submission information package is also beneficial in the long term as it can assumed that different tools will be used for harvesting data and create Submission Information Packages. Modified policies and new use cases may require additional. As a consequence the container formats as well as the amount and granularity of metadata may change in the future.

Content and metadata are normalized when the sSIP is generated from the SIP. The sSIP supports a single container format. All content data must be embedded in one or more WARC[6] containers. The British Library decided to use WARC as the standard container for web content since it became a NISO standard in 2009.

ARC containers which are provided by Heritrix must be migrated to WARC containers. The web content harvested by PANDAS need to be migrated into WARC containers as well. In this case, the HTML files need additional transformations as PANDAS modified all the

[4] http://www.loc.gov/standards/mets

[5] PREMIS Data Dictionary for Preservation Metadata, version 2.0, March 2008, http://www.loc.gov/standards/premis/v2/premis-2-0.pdf

[6] Web Archive File Format: http://www.digitalpreservation.gov/formats/fdd/fdd000236.shtml

hyperlinks in the harvested HTML files. Instead of keeping the original URL, the links are using relative URLs pointing to the appropriate files the local file system. As the local files are embedded in the WARC container, the URLs need be replaced by absolute URLs.

Normalizing data and metadata from two very different sources is a challenge when both sources provide a very different quality of data. It becomes even more difficult, when future sources, preservation actions and requirements need to be considered. New tools may provide additional information or transform content in different, yet unknown ways. For this reason data model must be easily extendable and flexible to accommodate additional metadata.

3.2. sSIP/AIP Content Model

The standardized Submission Information Package (sSIP) uses a similar structure than the AIP. Metadata as well as content are stored in the same way using the same standards for metadata and container formats.

Both information packages share the same underlying content model. The content model defines so called abstract entities. Abstract entities represent the objects containing which need to be preserved. Unlike the content model of the web crawlers, the data model for the sSIP and AIP disaggregates the preservation object from its digital manifestation.

Over time a number of different manifestations may occur. Content files will be migrated to new file formats. In abstract entity will remain and described by the same metadata record. Abstract entities are usually created by intellectual work. Therefore they are also called intellectual entities. Each of them may have a descriptive metadata record. This record describes the intellectual entity itself – e.g. the webpage and not its HTML manifestation.

The British Library's content for web archiving uses the following abstract entities:

- Website: a website is a collection of webpages which are interconnected and accessible under the same domain name. Usually the same organization or person is responsible for those pages. From the curatorial perspective a webpage must be regarded as preservation worthy in order to become part of a website. Not all webpages available under the same domain name become part of a website. A website must have basic descriptive metadata.
- Webpage: a webpage is a resource which is intended to be accessible and displayable as a distinct object. This resource is referenced by one or more hyperlinks which are forming the connections between webpages of a single website. Each webpage should have a descriptive metadata record. However in practice it proofed difficult to capture this metadata this metadata.
- Associated objects: An associated object is part of a webpage. Its rendition is embedded into the rendition of the whole webpage. The webpage provides the context for the associated object. An image being part of a webpage would be an associated object.

Every Webpage has a digital manifestation. It consists of at least one file or bytestream which can be interpreted and rendered to show the actual content of a Webpage. A digital manifestation may comprise several files. HTML based Webpages comprise of an html page, all referenced image files and Cascading Style Sheets containing important rendering information. An information system, such as a web browser, needs all those components to render a Webpage properly.

After a period of time this digital manifestation of a Webpage might become unrenderable. The manifestation or certain elements within the manifestation (e.g. the images) might not be rendered by common web browsers. The web as an open environment had to cope with incomplete support of standards as well as with different interpretation of standards from very early on. Complex Websites are often created for a certain group of browsers and browser versions. Once these browsers are not available anymore or unsupported by future operating systems, a migration might become a sensible alternative to provide further access to the content.

Whether a migration will result in a slightly different HTML file or a new file format (like PDF), it will create a new digital manifestation of a Webpage. Storing information about this migration process is essential for long term preservation. Preservation metadata attached to each file or bytestream must contain information about its origin. Providing an audit trail for a Webpage will ensure the authenticity of the data being stored.

A single manifestation consists of all content files which are required for a Webpage. All files are stored in WARC containers. For the initial crawl of a website a single WARC container will contain all files. The WARC container retains the curatorial coherence of the website.

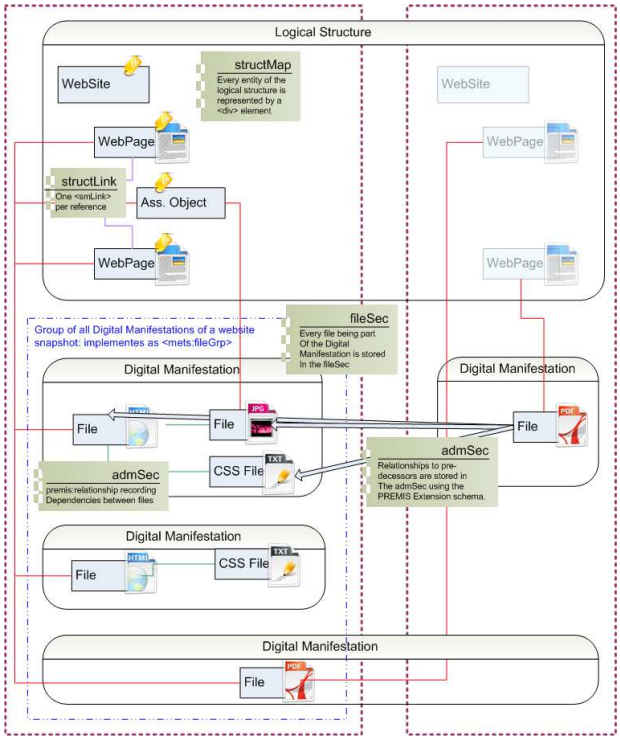


Figure 1. Content Model of the sSIP/AIP

The files of subsequent crawls or migrated Digital Manifestations may be stored across more than one WARC container. Every crawl will pass a deduplication process. This process detects any newly harvested bytestreams which had been stored earlier. In this case the file is not stored a second time. A migration might not impact all files. Some files of a digital manifestation might be unchanged. As no duplicates of files are stored, these unchanged files are stored in a separate WARC container than the migration results. In both cases the AIP's content is stored in several WARC containers.

3.3. Implementation of the Data Model

The data model was implemented in two stage. In the first stage the complete model was serialized using the METS framework. Every single object from the data model was represented by an appropriate XML element. As much metadata as possible were extracted, standardized and embedded into the METS descriptor including information from the logfiles, the WARC container and WCT database. The resulting METS descriptor for a single SIP/AIP is very large.

In the second phase the METS serialization was reviewed regarding storage size. The underlying data model wasn't modified, it's serialization was. The main aim was to reduce the size of the overall SIP. This lead to a few basic principles:

- Only those objects are defined in the METS about something needs to be said: either because appropriate metadata records are available or these objects need to referenced, accessed etc.
- Information is just stored once where possible; metadata is not stored in the METS file if it as already stored in additional files (WARC container, logfiles) which are part of the SIP/AIP.

The size of the METS container could be reduced significantly.

```
<mets:structMap TYPE="logical">
<!-- the  website containing  webpages -->
   <mets:div TYPE="WEBSITE">

      <!-- the first webpage -->
      <mets:div TYPE="WEBPAGE"/>

      <! definitios of image -->
      <mets:div TYPE="ASSOCIATEDOBJECT" />

      <!-- the second webpage -->
      <mets:div TYPE="WEBPAGE" />

  </mets:div>
</mets:structMap>
```

Figure 2. <structMap> in METS representing the logical structure of a harvested website

3.3.1. Structural Metadata

Abstract entities are represented by a element in the structMap section. The structMap section is the central section of each METS instance. Nested elements represent the logical structure of the website. According to the content model the hierarchy consists of three levels. For practical reasons the data model implements a simplified version of the content model. The data model's implementation restricts the hierarchical level of element to two. The uppermost elements represents the Website. All the other abstract entities are represented by elements which are direct children of the website's element. They are just available in cases there is something to say about the individual webpage or associated object (e.g. a metadata record is available). Otherwise the website's is the only in the structMap.

The content model defines relationships between webpages. They represent the navigational structure of a website and can be regarded as a site map. This information is also not recorded in the METS file for practical reasons. It would require parsing every html file to extract the links. Appropriate use cases justifying the additional effort are not regarded as relevant.

3.3.2. Descriptive Metadata

Every element may have one or more descriptive metadata records. Each metadata record is stored in its own <dmdSec> element. The British Library's web archiving profile supports Dublin Core and MODS records. All the descriptive metadata related to the website such as the title and subject are mapped to

MODS[7] using the latest 3.4 schema. Additional MODS elements such as typeOfResource, digitalOrigin and genre are set to fixed values. The Dublin Core elements being captured using the Web Curator Tool are recorded in a separate metadata section using the Dublin Core simple extension schema. Only those elements with very distinctive semantics are mapped to(e.g. title, creator) and recorded in the MODS record. Others with broad semantics such as dc:source are not mapable. These elements are only stored in the Dublin Core metadata section.

3.3.3. Rights Metadata

The web curator tool allows capturing basic rights metadata regarding the public access of data. This data is stored in the underlying relational database and used by the Access Tool when providing or restricting access to the content. As a consequence the rights metadata is proprietary. An extension schema had been derived from the WCT's database model. The rights metadata is recorded in its own administrative metadata section (<amdSec>) within the METS descriptor. Rights metadata are only available for the website. The administrative metadata section is attached to the element representing the website. Webpages or associated objects do not have their own metadata section containing access rights.

3.3.4. File Definitions

All container files are defined in the file section. Individual bytestreams within the container files are not defined unless there is a specific reason for it:

- A bytestream has a preservation metadata record which is needed in case of format migrations, recording certain provenance information etc.
- Deduplication: The content of a container belongs to more than one SIP/AIP; only relevant for the domain crawl, not for the selective crawl.

To distinguish the different purpose of bytestreams appropriate file groups are used. The web archiving profile supports the following groups:

- DigitalManifestation file group: The digital manifestation of all webpages and associated objects as well as their helper files (CSS, javascript etc.) are grouped into a single file group. It contains all files which are needed to render the whole website.
- Logfile file group: The Logfile group contains all logfiles. Created by the crawler (Heritrix or PANDAS). Logfiles provide useful provenance information. The crawl logfile contains information about every URL which had been requested. Error logfiles track any error which occurred during the harvesting process and may indicate the reason why a bytestream is not available in the AIP/SIP.
- Viral Files: all infected files are defined in this group; they are not regarded as part of the digital manifestation.

```
<mets:mets>
   <mets:fileSec>

      <!--
    The Digital Manifestation file group
     with the definition of two files -->
      <mets:fileGrp
              USE="DigitalManifestation">

      </mets:fileGrp>
      <!-- define a separate group for
            logfiles -->

      <mets:fileGrp USE="Logfile">
      </mets:fileGrp>

      < mets:fileGrp USE="ViralFiles">
      </mets:fileGrp>

   </mets:fileSec>
</mets:mets>
```

Figure 3. <fileSec> in METS defining different file groups.

As mentioned above all the data is stored in WARC containers. A WARC container consists of a so called WARC records - one for every successful http-request. Beside the actual content data it stores information on the http-protocol level such as response codes, file size and format type. Every WARC record is compressed within the WARC container.

In case a single bytestream needs to be defined in the METS, the complex structure of a WARC container must be represented. The container, the individual WARC record and the bytestream within the record are all represented by nested <file> elements.

The WARC record's <file> element contains specific information about the location of the WARC record within the container. The appropriate byte offsets are stored in the <file> element's BEGIN and END attributes. These two attributes were just introduced into the METS schema since version 1.9.

The same mechanism is used for recording the start and end of the actual content within the WARC container is stored in the content file's <file> element using the BEGIN and END attributes as well. It is important to note that before the actual content can be retrieved from the WARC record it needs to be uncompressed first. The <transformFile> element indicates which algorithm must be used for uncompressing the WARC record.

The value for the BEGIN and END attributes are also stored in the proprietary CDX files created by the Heritrix crawler. These files are index files for a WARC container and are used for randomly access content

[7] http://www.loc.gov/standards/mods

bytestreams from the WARC file. They are not part of the Archival Information Package as they are regarded as an access file. Its information can easily be reconstructed from the WARC container itself.

```
<!-- the WARC container itself -->
<file ID="container01">
  <transformFile
  TRANSFORMTYPE="decompression"
  TRANSFORMALGORITHM="WARC"
  TRANSFORMORDER="1"/>
   <!-- the WARC record within the WARC
        container -->
  file ID="gzip01" BETYPE="BYTE"
       BEGIN="20" END="22674">
    <transformFile
    TRANSFORMTYPE="decompression"
    TRANSFORMALGORITHM="GZIP"
    TRANSFORMORDER="1"/>

       <!-- the content bytestream within
          the WARC record -->

       <file ID="contentfile01"
       BETYPE="BYTE" BEGIN="623" END="35143"
       CHECKSUM="xxxxxx"
       CHECKSUMTYPE="SHA-512" SIZE="35123"
       MIMETYPE="text/html"/>
```

Figure 4. Example showing METS structure recording the internal structure of a WARC container file.

The first version of the METS profile defined every bytestream as a <file> element. The idea was to support (future) end-to-end business processes embedding all necessary information in the AIP. But as the current dissemination tool (wayback machine) doesn't support METS and the byte offset information can easily be extracted from the WARC file itself, the review regards the index information as redundant and consequently abandoned it from the SIP/AIP.

But having defined the mechanism for storing this information in METS, the AIP could record and provide all necessary information which is required for the dissemination of content. The data-requirements of the Access Tool had been considered when defining the AIP's data structure.

In case a bytestream is represented by a <file> element it must have an administrative metadata record attached. It contains basic preservation metadata as well as digital provenance information.

3.4. Preservation Metadata

The British Library's web archiving profile uses the PREMIS metadata schema as an extension schema to METS. PREMIS records are stored within each file's administrative metadata section (<admSec>). Content files, helper files and container files must have an administrative metadata section. Though WARC records are represented by a <file> element they don't have a metadata record of its own.

3.4.1. Technical Metadata

The preservation metadata record stores basic technical information about each file:

- Checksum: the SHA-512 checksum is calculated and recorded in the <premis:messageDigest> element as well as in the CHECKSUM attribute of the METS' <file> element.
- Size: The <premis:size> element records the size of the content bytestream in bytes. It contains the same information as the SIZE attribute of the METS' <file> element.
- Original URL: the URL which had been used in the http request. This URL is hostname based and may therefore not specify the actual server which submitted the bytestream to the crawler. In load balancing and virtual server environments the http-request may be redirected internally. The hostname based URL is recorded in the <premis:originalName> element and is retrieved from the crawler's logfile.
- File Format: The file format is retrieved from the HTTP-header in the http-response as it is recorded in the crawl log. The file format information is extracted from the crawl log and captured in <premis:format>. For the AIP this information is enriched with an appropriate reference to the PRONOM file format database using the DROID tool.

The METS descriptor stores preservation metadata for the container file as well. As the WARC file is assembled during the crawling process is does not have an original filename. The format information is set to "application/warc" as this is the official MIME type of the file format.

3.4.2. Provenance Metadata

Provenance metadata is recording the history of a digital object. PREMIS provides an event framework for storing events within the bytestream's preservation metadata record. During a bytestream's lifecycle various events will have an impact on the object. Most events are occurring as part of well defined business processes. The METS profile defines all the events which may occur during the end-to-end process and have an impact on the web content during its life cycle. As business processes may change in the future, the event model may be extended with additional events.

Some events are only extracting or verifying information and don't have any impact on the bytestream itself. Other events have an impact on the bytestream as they are creating or modifying the content itself. To provide a comprehensive audit trail and ensure the information's authenticity it is important to record all events. Each event has a timestamp, an outcome and an associated agent. PREMIS defines an agent as a separate

entity. Agents may be persons or software systems which were responsible for an event.

```
<premis:event>

    <premis:eventIdentifier>

      <premis:eventIdentifierType>local
       </premis:eventIdentifierType>

     <premis:eventIdentifierValue>event01
       </premis:eventIdentifierValue>

    </premis:eventIdentifier>

    <premis:eventType>migration
     </premis:eventType>

    <premis:eventDateTime>2006-07-16T19:20:30
    </premis:eventDateTime>

    <premis:linkingAgentIdentifier>

     <premis:linkingAgentIdentifierType>local
     </premis:linkingAgentIdentifierType>

    <premis:linkingAgentIdentifierValue>
      agent001
     </premis:linkingAgentIdentifierValue>

    </premis:linkingAgentIdentifier>

</premis:event>
```

Figure 5. Representation of an event in PREMIS

The web archiving profile defines two different events: virus check- and migration event. The harvest process is not recorded in the PREMIS record as metadata should not be stored redundantly. The logfiles whicha re part of the SIP and AIP are already containing appropriate provenance information such as the URL, IP-address, datetime stamps and the http-return code.

Virus-Check Event: Each file which is ingested into the archival store is checked for viruses. The virus check event is recorded on level of the WARC file. Only in the exception that a virus had been detected, the appropriate bytestream is defined in the file section and an appropriate PREMIS record with the event information is recorded on bytestream level.

In case a virus is detected the ingest system tries to clean the effected bytestream. This may or may not be successful. Depending on the success, the output of the event is recorded in the <premis:eventOutcome> element:

Value for eventOutcome	*Virus check outcome*
no virus detected	No virus detected
viral, cleaned	virus detected, bytestream had been cleaned successfully
Viral, failed but forced	virus detected, but cannot be cleaned

Table 1. Event outcome values for the virus check event

The <premis:agent> element records the anti-virus software and its virus database version which had been used during this event

In case the viral file could be cleaned, the original, viral file is stored in a separate file group marked as “viralfiles”. It is not part of the website’s digital manifestation. The viral file is only kept for administrative purposes and will not be accessible by end users via the Access Tool. Instead the cleaned file will be part of the website’s digital manifestation file group.

To keep track of the bytestream’s history the relationship between the cleaned and infected bytestream is kept in its preservation metadata record. The <premis:relationship> element records a pointer to the old viral bytestream. The relationship type is set to “derivation” and its subtype to “cleanedFile”.

Transformation Event: HTML bytestreams which had been harvested using PANDAS need to be transformed. All URLs need to be updated. This transformation process takes places prior to ingest into the Archival Store. The appropriate transformation event is recorded as part of the standardized Submission Information Package as well as in the Archival Information Package.

Rewriting the URLs results in a set of new HTML files. Though both sets of HTML bytestreams are defined in the METS descriptor, only the new HTML bytestreams are part of the Digital Manifestation. The original bytestreams are part of a separate file group. Image bytestreams, style sheets etc. are just part of the same Digital Manifestation as the new, transformed bytestreams.

The transformation and the relationship between the old and new bytestream are recorded in the ew btyestream’s preservation metadata record .

Element name	*value*
premis:relationshipType	Derivation
premis:relationshipSubType	Transformation

Table 2. Relationship between PANDAS html files and transformed html files.

A transformation is regarded as successful, whenever the new bytestream exists. The <premis:eventOutcome> Element can only contain the value “success” for the transformation event.

Migration Event: The Archival Store supports WARC as the only container format for web content. All ARC containers are migrated into WARC containers prior to ingest. This migration process is described in the WARC-container’s preservation metadata record.

The outcome of this event always “success”; otherwise the WARC container would not exist. A relationship between the WARC and the ARC file is described using the <premis:relationship> element

pointing from the WARC file's to the ARC file's preservation metadata record.

Element name	*value*
premis:relationshipType	Derivation
premis:relationshipSubType	Migration

Table 3. Relationship between WARC and ARC container files

When container files are migrated, the actual content bytestreams stay untouched. Consequently event information for individual bytestreams is not recorded.

4. CONCLUSION

The three different information packages which are defined by the OAIS are used for three very different purposes. Though the British Library's METS profile for Web Archiving does not define a Dissemination Information Package, it supports the end-to-end business process of harvesting, ingesting and preserving web pages and enabling long term accessibility. The METS profile considers access as well as preservation requirements.

The practical implementation does not make use of the whole complex data structure: METS files become fairly large and a lack of support of METS by dissemination tools makes it inconvenient and expensive to store all the metadata redundantly. Instead the profile ensures that all the metadata is part of the SIP/AIP – either as part of the METS descriptor itself, as part of a proprietary file (logfiles) or embedded in and restorable from the actual content file (index of WARC files).

Using standardized metadata frameworks and schemas such as METS and PREMIS are as important as an extendable and flexible content model. Defining abstract entities and their manifestation as separate objects allows future implementations of tools to support a complex end-to-end process without relying on proprietary data structures.

It ensures easy integration of the web archiving content with other content streams and library systems. As the operational and archival metadata is now being managed in the library's preservation repository content can actively be preserved. Preservation actions can be carried out, new digital manifestations of web content can be created.

FULL PAPER

RELEVANT METADATA TO PRESERVE "ALIEN" AIP

Angela Di Iorio **Maurizio Lunghi**
Fondazione Rinascimento Digitale
Via Maurizio Bufalini Firenze

ABSTRACT

This article describes the development of Archives Ready To Archival Information Packages (AIP) Transmission a PREMIS Based Project (ARTAT). Following the project approach, the starting phase consisted of prototyping a layer conveying preservation metadata, which can be encoded from the existing archival systems, and exchanged with other repositories. This layer called Preservation Metadata Layer (PML) uses PREMIS semantics as the common language to overcome archival systems differences, and to transmit out of its original context, relevant preservation information about content objects comprising an AIP. Since a repository, following the OAIS reference model, usually provides resources with metadata container objects, the experiment performed an analysis on commonly used container formats, in order to enable the traceability of semantics from a local to extra-local level, and the technological understandability of alien AIPs. The analysis has allowed the definition of a PML data model, laying the production of prototypes. The adoption of common semantics, like PREMIS, supports the opportunity of preserving correctly alien AIPs, coming from different technological environments, and hopefully enables the overcoming of obstacles to the interoperability among diverse archival systems.

1. INTRODUCTION

This article describes the development of the project named Archives Ready To AIP Transmission a PREMIS Based Project (ARTAT) [3], that took place, from March to April 2010. The goal of ARTAT is to experiment with the adoption of a common preservation metadata standard as an interchange language in a network of cooperating organizations that need to exchange digital resources with the mutual objective of preserving them in the long term. The project in pursuing its initial objectives, has experimented with the definition of a Preservation Metadata Layer (PML) following the PREMIS standard Data Dictionary (DD) specifications [8] that will integrate repositories' preservation metadata. The exported repositories' AIPs [2] including a PML will be received by selected repositories and ingested into their archival systems. Hopefully, because of the common PREMIS knowledge base, the receiving repositories will be able to locate information objects and data objects contained in the AIPs transmitted by the originating repositories.

To date, the project consisted of testing the Preservation Metadata Layer prototype produced from representative samples, selected from the initial participant repositories.

The critical path analysis, which was conducted on the PML prototypes, will be traced in order to support the ultimate objective of transmitting resources destined for preservation in a repository other than the originating repository. It is assumed that the devised layer will be agnostic about the originating archival systems, as well as about the receiving archival systems. A successful transmission can be accomplished as long as both of the repositories in the transfer can manage XML conforming to the PREMIS framework.

The milestones that will be explained below aim to test the feasibility of inventing a layer which contains all relevant information for the receiving repository to offer long term preservation services in the foreseeable future.

More information about aim, objectives, tools and methodologies of the project are available on the Fondazione Rinascimento Digitale website which supported this project and the Italian PREMIS community (http://www.rinascimento-digitale.it/projects-artat.phtml).

2. ARTAT PROJECT OVERVIEW

The ARTAT[1] project started in March 2010 with interviews conducted with the first three participants:

- ICCU's MAGTECA[2] an institutional repository which collects resources from geographically dispersed Italian cultural heritage institutions

1 http://www.rinascimento-digitale.it/projects-artat.phtml
2 www.internetculturale.it

- Magazzini Digitali (MD)[1] a project undertaken by Fondazione Rinascimento Digitale and National Library of Florence to preserve Italian doctoral theses for the long term
- The digital repository of the Library & Archive of the British School at Rome[2]

The interviews followed the inquiry phase that were reported in the project workplan.

The project aims to provide existing digital repositories with a layer of preservation metadata that is exchangeable with other repositories. The focus is not on changing existing archival systems, but rather on creating the ideal conditions for exchanging resources, strengthening their own management with a view to long term preservation, and enabling opportunities for offering preservation services to third parties.

The experiment's approach is to define and test a preservation metadata layer, encoded according to the PREMIS standard.

The export of a repository's AIPs with a PML provided should enable selected repositories to receive and ingest them into their own repository systems.

The building of PML will originate from the archival management system and, through a controlled data flow, will feed the PML exchanged with other receiving repositories.

The approach taken from the beginning of the project has two phases: the first inquiry phase, where participants are interviewed about their repositories' architectures and their management of preservation metadata, and the second PML production phase, which experiments with the translation of metadata contained in AIPs into the PML layer encoded in PREMIS semantics.

3. INQUIRY PHASE RESULTS

3.1. General consideration about initial application of the questionnaires

The initial inquiry phase was conducted in March 2010 with the initial participating repositories and concluded at the beginning of April 2010. The information was obtained through interviews generally guided by semi structured questionnaires. The initial questionnaires, which focused on archival systems and preservation metadata management, were dramatically reduced during the interviews, because it became clear that in spite of all technological differences, nearly all systems mainly contain metadata useful to the preservation but they do not manage it as preservation metadata. A side effect of the interviews was to make repositories' managers aware of the risks of the lack of management of preservation metadata.

The results of the questionnaire as well as the report of information gathered during the prototyping will be published on the ARTAT website for the preservation community.

A review of the questionnaire will be conducted and submitted to the future project's partners.

3.2. Repositories technologies overview

The information gathered from the inquiry phase regarding the metadata schemas managed is summarized in Table 1. This is the basis from which we have started to address the problem of differences in standards adoption, as well as to find a solution in overcoming the interoperability issues that in practice limit the understandability of AIPs, exchanged by repositories.

Knowledge about the metadata container standard adopted by the repositories is an important starting point of the experiment. Analyzing the application and composition of the containers used, and the comprehensiveness of information gathered inside, is important in order to structure correctly the PML description.

Institution/ Project	Metadata type	XML Schema name	Version
ICCU	Container	MAG	1.0-2.01
	Descriptive	DC simple	1.1
	Technical	MIX	0.1 draft
MD	Container	MPEG21-DIDL	-
	Descriptive	DC simple	1.1
	Technical	Jhove	1.5
	Technical	MIX	0.2
BSR	Container	METS	1.9
	Descriptive	MODS	3.3
	Descriptive	DC simple	1.1
	Technical	MIX	2.0

Table 1. Metadata schemas used by the interviewed repositories

The evidence of semantics adopted in metadata containers is useful to the likely exchange scenario, allowing the data conversion, from the repositories internal structure to PREMIS [5].

At the end of the repository's inquiry phase, repository managers were asked to submit a sample metadata object encoded in XML that is representative of their AIPs.

4. THE PRESERVATION METADATA LAYER (PML)

4.1. PML Target

The target of preservation is the information package defined as the AIP in the OAIS conceptual model. This package, in actual applications, consists of content and metadata. Without consideration for how the AIP is managed by the archival systems or whether the metadata encoded in XML is used to support one or more OAIS process (submission, archival,

1 http://www.rinascimento-digitale.it/magazzinidigitali.phtml
2 http://digitalcollections.bsrome.it/

dissemination), the focus of PML is the XML metadata files. In particular, the PML target is all files that package different metadata categories together in a formally declared structure, and that usually are defined as metadata containers, like for example the METS files. More specifically "A *container* is the unit for aggregating the typed metadata sets, which are known as *packages"* [7]. In ARTAT approach these files will be considered as objects, conforming to the PREMIS data model specifications. As Metadata Container Object (MCO) is meant the container file object that can bind different types of metadata objects and content objects together, by means of the embedding or referencing mechanism.

The MCO samples coming from the participating archival systems will be submitted to the PML prototyping process. The outcomes will confirm the feasibility of translating the system internal AIP into an "exchangeable AIP", which in ARTAT terms, means an AIP provided with a PML.

The PML is essentially a translation of the content and the relationships among the constituent objects (metadata and content) of an AIP.

4.1.1. Metadata Container Objects

Usually, metadata containers are used to package different types of metadata and can fulfill different OAIS functions. The interoperability difficulties that arise when containers are used in contexts outside of their original archival systems are well known. These difficulties are caused by differences in structural design and in different levels of granularity of metadata application.

An MCO can be used by repositories to support the various functions specified in the OAIS conceptual model [2]. An MCO is a composite that can contain a diverse set of structured information conforming with formally specified semantics. As such, it is a purpose-specific object type. Usually, metadata containers are intended to bundle various types of metadata that describe the resource from different points of view. For example, METS is a widely used XML container format that wraps metadata types in well-circumscribed sections. METS can contain information about objects, both content and metadata, which is embedded (mdWrap) or referenced (mdRef) in some way.

Finally, to support their implementation, container standards have bindings in XML schema which may organize information quite differently from its original structure.

The characteristics of MCOs in use can be a significant factor when an exchange involves different MCO standards. Consequently, exchange packages derived from local MCOs, need to be structured with a common and well defined set of information, overcoming the local coding practices and constraints [6].

4.2. PML Structure

In ARTAT, a Preservation Metadata Layer (PML) will be added to the AIP by the originating repository, which needs to translate its native information into PREMIS semantic units. PREMIS was not originally designed to be a transmission format, but in ARTAT it is used to exploit the knowledge base focused on preservation metadata and founded on a well known model.

The PML is composed of two parts: the PML core and the PML redundant part, which together describe technically and structurally the AIPs content.

The PML core is the part which essentially translates the container's relevant metadata into PREMIS semantic units. The translation consists of a mapping from the original administrative, technical, provenance, rights and structural information into the PREMIS framework.

The PML redundant part simply describes the content objects in PREMIS terms mapping information like *objectidentifier*, *compositionlevel*, *fixity*, *size*, *format*, *originalName*, and *storage* from the object's related metadata.

The PML consists of one or more PREMIS files connected by internal and external identifiers, and connected by reference, to the AIPs' metadata and content objects.

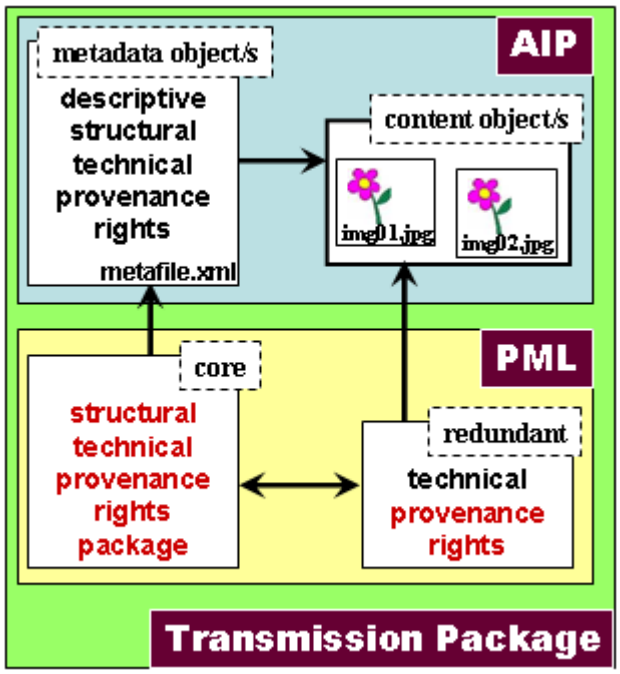


Figure 1. Transmission Package structure composed of Archival Information Package and Preservation Metadata Layer.

4.3. PML Coding and requirements

The PREMIS metadata standard was selected for the PML because it is strictly focused on preservation metadata and because it has been widely implemented in the international preservation community. The choice

was made on the assumption that the standard is built on well-defined semantics and a well-known data model, so it ought to be conducive to interoperability at organizational and technological levels.

The ARTAT project defined three main requirements for the PML. The first requirement is PREMIS conformance, which requires: following the specifications of PREMIS Data Dictionary names and definitions for semantic units, adhering to Data Dictionary applicability guidelines, conforming to repeatability and obligation stipulations, and using mandatory semantic units as the minimum amount of metadata useful to preserve digital objects in the long-term. The second requirement is to provide PREMIS metadata as comprehensively as possible, in order to facilitate the receiving repository correctly understanding the PML, since the originating repository could have some missing or implicit preservation metadata. The third requirement is the independence of the PML from the AIPs, making its reuse easier and its preservation feasible in different technological contexts.

4.4. PML application context

The cooperative context held by the agreement among different partners that manage diverse archival systems is the ideal application context where AIPs can be exchanged in order to share the preservation responsibility or also to provide or receive third party preservation services.

In this context the project predicted the transmission scenario (par.4.5) where AIPs are provided with PML by the originating repository which makes the "translation in" PREMIS code. The whole package, AIP and PML, is transmitted to the receiving repository system which acquires and "translates out" the PML and archives the objects as its own AIP.

The transmission package is the set of the XML formatted original AIP (content objects and metadata objects) and the preservation layer as PML (core and redundant) which is the translation part understandable by the different systems.

The cooperative context will be supported as much as possible by the adoption of common controlled vocabularies in order to translate the PML. The adoption of controlled vocabularies, as well as shareable nomenclature systems, for example the agent information, will facilitate the automatically encoding of the precompiled set of the PREMIS semantic units.

4.5. PML Transmission scenario

The AIP with metadata translated into a PML will constitute the transmission package. The transmission will happen in some formally established way, where the agreements' terms will be explored in further investigations.

The originating repository *A* which holds the AIPs performs the PML "translation in". The receiving *B* repository performs a PML "translation out", which consists of reading PML core metadata, detecting the MCO structure, the AIP's metadata, and the content objects and their relationships. Finally *B* pieces together the PML jigsaw, interpreting the original AIP and creating a new *B* MCO corresponding to *A*'s MCO in its own archival system, which will manage all of the original AIP objects plus the original *A* MCO. *B* MCO is connected to *A* MCO by means of the digital provenance information (events and agents) and objects' relationships.

This mechanism was devised to avoid the loss of information, which is natural when you make a mapping from one standard to another.

In the envisaged context, the receiving repository will act on the alien AIP just to complete the migration and to preserve its integrity and authenticity or to perform other predetermined preservation actions. In this scenario, it is not supposed to make any modification of original AIP or MCO but only integrate the AIP. All the events that have affected the AIP's objects will be recorded in some way as *B* MCO conforming to the PREMIS Data Dictionary specifications.

Two other possible transmission scenarios are:

- transmission back to the originating repository *A*: the transmission is performed in the same way, *B* makes a PML "translation in" of its AIP. The resulting PML should contain the same structure as the former transmission *A* original AIP and MCO, plus *B* MCO integrated with the events that occurred in the elapsed time. *A* makes a PML "translation out" of PREMIS information differences, which occurred in managing or updating actions;
- transmission forward to other receiving repositories *C, D…*: *B* MCO with relative AIP (objects plus original *A* MCO) is translated in PML; *C* translates out *B* MCO recording the *C* MCO digital provenance (from *B* MCO, from *A* MCO); *C* translates in its AIP, *D* translates out *C* MCO recording the *D* MCO digital provenance (from *C* MCO, from *B* MCO, from *A* MCO) and so on.

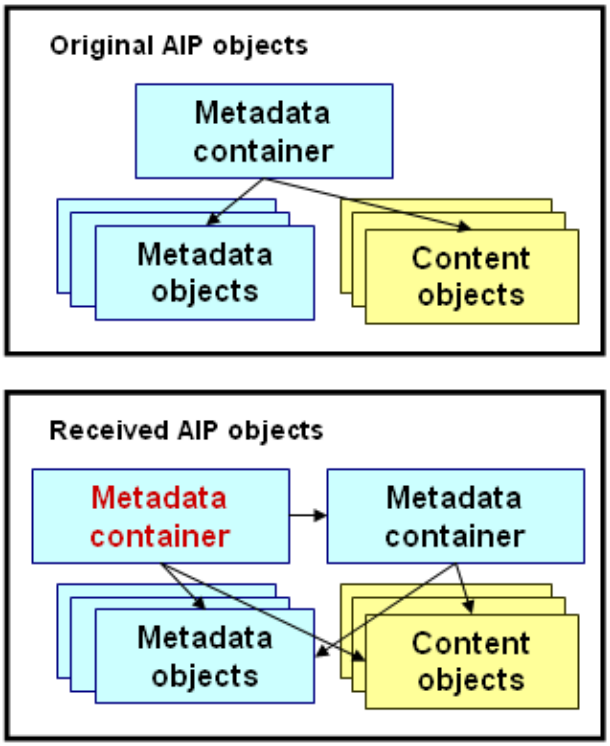


Figure 2. Differences in transmitting repositories of the Archival Information Packages.

5. METADATA CONTAINER OBJECTS ANALYSIS

The PML prototypes built from the sample files of metadata objects obtained from the participating repositories, were realized through the following milestones.

5.1. Samples' analysis process

An analysis of samples was performed in order to verify the existence of all necessary elements for building the PML encoded in PREMIS and to comply with requirements.

The sample files are encoded in three different metadata containers: the most common in the digital library community, METS[1]; the multimedia framework MPEG21-DIDL[2]; and the Italian application profile MAG[3].

Despite the containers' differences in the information framework architecture, the samples analyzed contain at least one descriptive section well circumscribed. The structural metadata are gathered in a formally defined section or in hierarchical elements, structurally added.

The administrative information usually is scattered in different sets that can be delimited in a fragmented way as technical, provenance, or rights. Despite the fragmentation, the presence of these sets of metadata should be considered obligatory in transmission contexts, even though the MCO XML schema doesn't require them as mandatory. Conforming to the obligation rules declared in the MCO schemas, a METS document can have only one structural section, MAG can have descriptive and only some of administrative metadata and in MPEG21-DIDL it is sufficient to declare only a didl:Item element. This is obviously not sufficient to describe a digital resource from a preservation point of view, but actually the repositories use containers in a sufficiently exhaustive way to describe their resources.

5.2. Samples' analysis results

Considering lessons learned in the transfer context of TIPR (par.6), and the necessary maintenance of metadata quality at a non-local level, the analysis has detected the existence of the mandatory PREMIS DD semantic units as well as the lack of or the inefficiency of information at a cooperative level. The following list is a draft of the information areas where ARTAT has to make metadata integration in order to cover cooperative needs:

- the object's identifier system, has to be refined and customized in order to identify unambiguously objects, agents, events in a nomenclature system recognizable by all ARTAT partners;
- the rights declared into three samples referred to access conditions for the resource as whole. The METS samples, the copyright information was replicated in both the descriptive section and in the METS rights section. The rights in the PML core will cover the rights and permissions about the transmission package since more detailed rights and permissions applied to the single objects will be replicated into the PML redundant. A shareable rights framework system has to be developed in order to supply the needs around third party preservation;
- events information is managed by archival systems but are not yet implemented in the MCO consequently events semantic units will be integrated at the first provision of the PML;
- the agents are not provided homogeneously but will be added automatically from the partners' nomenclature system.

6. LESSONS LEARNED FROM TIPR PROJECT

The goal of the Toward Interoperable Preservation Repositories (TIPR) [1] project is to experiment with the transfer of complex digital objects between dissimilar preservation repositories that need to be able

[1] http://www.loc.gov/standards/mets/
[2] http://mpeg.chiariglione.org/standards/mpeg-21/mpeg-21.htm
[3] http://www.iccu.sbn.it/genera.jsp?id=267

to exchange copies of AIPs with each other. The ARTAT project has similar objectives but the application context is slightly different, because it cannot rely on the knowledge base of a single container format like METS. For this reason the PREMIS translation methodology has been adopted to overcome the interoperability issues due to the differences in the container adoption.

The development of ARTAT has taken into account the issues and the outcomes obtained by the transfer test of TIPR outlined in the referred article [1]. The TIPR requirements are: 1) based upon METS and PREMIS, 2) exchange package flexible, agnostic about the internal structure of AIPs, 3) exhaustive at package and representation level, 4) selected information must be understood by the receiving repository.

The TIPR approach is to define a common exchange package format, the Repository eXchange Package (RXP) where certain information critical to digital preservation must be, not only stored, but also understood based on the concept that a meaningful exchange can be achieved with semantic interoperability.

The information gaps that emerged from TIPR transfer tests results and ARTAT lessons learnt are:

- TIPR found information pertaining to the exchange package (history, description, and high level rights) must at this time be recorded at the intellectual entity level, because the highest level of object describable in PREMIS is a representation object. The PML core gathers events and rights at the exchange package level;
- both TIPR and ARTAT found problems with the unambiguous identification of entities;
- details about RXP composition by the source repository – relationships' information of PML core;
- how a packages will be transferred from source to target repository - devising partnership's agreement and transmission conditions applicable to the massive transmission of AIPs;
- actions to be performed - providing a common controlled vocabulary about actions that must be selected at PML production time and associated with agents;
- rights and permissions - rights framework system;
- archiving and preservation treatment - partnership's agreement level;
- financial and legal aspects of agreement - should be provided in ARTAT partnership agreement.

These lessons learned have affected the following PML data model.

7. PML DATA MODEL

The PML data modelling milestone consists of a selection of metadata elements from the PREMIS DD.

The data model in this context can be defined also as an obligation model, because it summarizes the mandatory elements necessary for AIP transmission.

Conforming to the PREMIS DD specifications, the mandatory semantic units pertaining to objects will be obligatorily used for the PML core and for every object's information, and replicated into the PML redundant part: *objectIdentifier*, *objectCategory*, *objectCharacteristics*, *storage*.

In the *objectCharacteristics* container, the optional semantic units *fixity* and *size* are considered mandatory for AIPs transmission, in the cooperative preservation context. These semantic units are considered useful because they allow the receiving repository to compare the original objects characteristics information to that processed by its own archiving system on the translated AIP.

Even though the actual prototyping did not use the digital signatures, this PREMIS metadata container might be considered mandatory for future transmission tests, to support the assessment process of the origin and the integrity of packages transmitted.

The semantic units pertaining to Agents are considered mandatory to identify the originating repository, as well as the receiving repository, in order to trace the chain of responsibility. All agents' semantic units will be supplied automatically, thanks to the ARTAT partners' nomenclature system.

The semantic units pertaining to Events *(eventidentifier*, *eventype*, *eventDateTime eventDetail*, *eventOutcomeInformation)* are all mandatory to describe the event history of the objects. The first version of the PML will include events' records will be produced, detailing this operation. Further events information should be provided if existing systems are integrated with events management functions.

Considering the transmission objective, the PML rights at PML core level will include the following semantic units: *rightbasis* (by default a "license" where all terms of the agreement are defined), *licenseInformation* which specifies metadata about license document and *rightsGranted* which specifies the actions that receiving repository can perform on AIPs.

The PML data model design and the anticipated transmission scenario, led the project to the early belief that significant properties and relationships are critical for conveying the structure of AIPs. The particular role played by these elements, will require more tasks focused on ascertaining the correct communication of the AIP's internal structure.

7.1. Significant properties of metadata container objects

Since the target of the PML core is the MCO, the actual literature about characterization of digital objects was consulted in order to identify the significant properties

of the MCO. The latest outcomes from the INSPECT[1] project, which gathered and leveraged all the former projects on this topic like CEDARS[2], CAMILEON[3], DELOS[4], CASPAR[5], PLANETS[6] etc., were found to be extremely useful.

As defined by the INSPECT project significant properties are "The characteristics of digital objects that must be preserved over time in order to ensure the continued accessibility, usability, and meaning of the objects, and their capacity to be accepted as evidence of what they purport to record". It is evident that MCOs are themselves digital objects that encompass all relevant information needed to make referred objects accessible, meaningful, authentic and reliable. In our context, where information has to be not only transmitted but also properly interpreted from other systems, it would be useful to subject the MCO to the INSPECT workflow analysis [4], in order to convey the significant properties to third parties.

The workflow consists of three sets of activities: Objects analysis, Stakeholder analysis, and Reformulation. MCO analysis and experiment will be detailed in the coming months, but a draft of the ongoing activities of this task is showed in Figure 3.

Some of the steps of objects analysis are listed here:

- *Identify the purpose of technical properties:* Considering the INSPECT categories, the *content* of MCO is XML text; the *context* is the environment, where the participants manage metadata and its exchange; the *rendering* is considered the recreation of an AIP in a recipient repository by means of a translated MCO, where metadata values and relationships among metadata objects and content objects are replicated in a new container; the *structure* is metadata which contains information about intra-relationships and inter-relationships; the *behaviour* is how the information object is connected to other metadata or content objects (i.e. the mdRef for external metadata files used in METS).
- *determine expected behaviours:* Limiting the analysis to the transmission context, where a source and a recipient have to exchange AIPs between their heterogeneous archival systems, the stakeholders involved in transmission of AIPs are repositories' systems that have to be able to make an interpretation of the alien AIPs and to ingest them as their own AIPs. This particular "user" with a well defined objective may wish to perform the following main activities: selecting information relevant to preservation, interpreting technically the selected information, and understanding the relational structure conveyed.

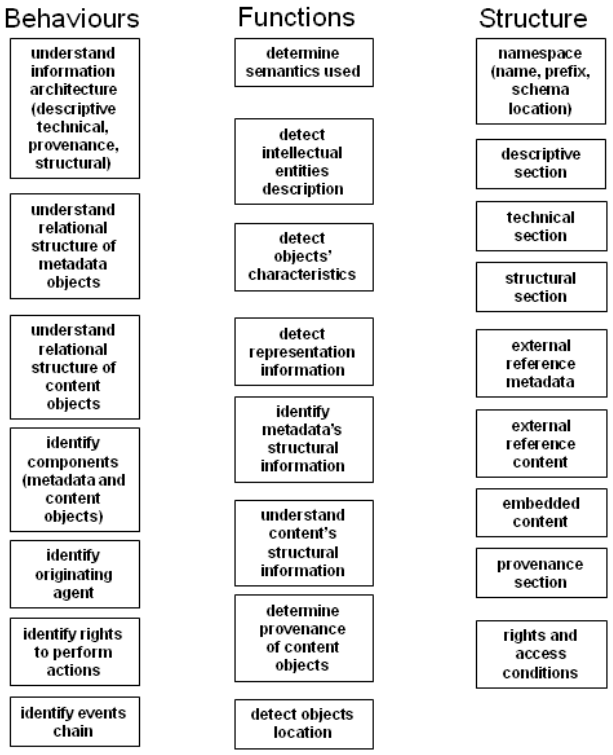


Figure 3. Draft of INSPECT workflow for Metadata Container Objects.

The premise underlaying the future experiment on the MCO are the following.

The hypothetical MCO should contain information about the schema used, validation outcome, authenticity, complex inter-relationships with other metadata container objects and intra-relationships with content objects and other metadata objects (i.e. technical metadata externally referred).

Furthermore, the need to determine two types of information has been recognized: 1) information created by the originating repository that is intended to transmit to the receiving repository; 2) information establishing the provenance of an AIP indicating its purpose and the processes through which it was created and transmitted.

The authenticity and integrity of the MCO has to be maintained, in order to demonstrate that the MCO exchanged is what it purports to be. Consequently the identification of the originating repository as well as the receiving repository/repositories are important information, because the MCO is used for a specific purpose. Also the digital provenance information guarantees the continued authenticity in the future.

The experiment has not yet been in practice performed, but will consist of the production of MCOs encoded in different metadata container standards. The MCOs will be submitted to the related participating repositories that manage the same container format, in

[1] http://www.significantproperties.org.uk/
[2] http://www.ukoln.ac.uk/metadata/cedars/papers/aiw02/
[3] http://www2.si.umich.edu//CAMILEON/
[4] http://www.delos.info/
[5] http://www.casparpreserves.eu/
[6] http://www.planets-project.eu

order to test the feasibility of translation, and the exhaustiveness of significant properties as determined by the applied INSPECT framework analysis.

The outcomes of submission to repositories of the proposed MCO, resulting from the INSPECT analysis, will drive the revision of PML data model.

7.2. AIP's Relationships modelling

The MCO intra and inter-relationships with content and metadata objects will be described by means of a structured set of information. The *relationshipType* semantic unit has been defined for recording the conceptual connection among pieces of information: descriptive, structural, technical, provenance and rights.

In addition the value "referencing" was taken into account for outlining the simple reference to a content object or a metadata object.

At this time the following values have been defined for *relationshipSubType*:

relationSubType
external metadata/content
internal metadata/content
metadata wrapper

The Figure 4 shows graphically how the relationships [9] between metadata and content objects can be defined.

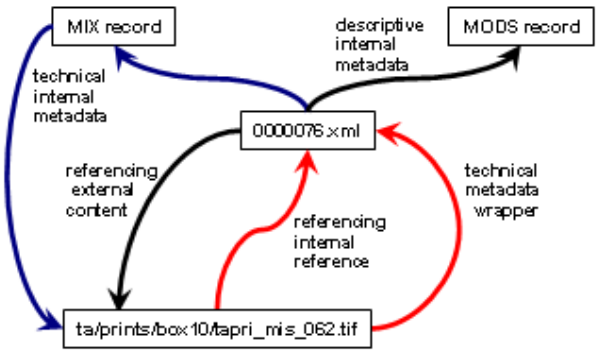


Figure 4. Relationships' prototyping.

To provide more information about the metadata schema used, the prefix and version has been tentatively added, but it is supposed that this information is related in some way to the significant properties of the MCO.

The figure below shows the corresponding simplified PREMIS code for the PML core and PML redundant.

PML core

objectIdentifier: 0000076.xml
relationshipType: descriptive
relationshipSubType:internal metadata:MODS 3.3
relatedObjectIdentification: MODS0000076
relationshipType: technical
relationshipSubType: internal metadata:MIX 2.0
relatedObjectIdentification: tif-138
relatedObjectIdentification:
ta/prints/box10/tapri_mis_062.tif
relationshipType: referencing
relationSubType: external content
relatedObjectIdentification:
ta/prints/box10/tapri_mis_062.tif

PML redundant

objectIdentifier: ta/prints/box10/tapri_mis_062.tif
relationshipType: referencing
relationshipSubType: internal reference
relatedObjectIdentification: 0000076.xml
objectIdentifier: ta/prints/box10/tapri_mis_062.tif
relationshipType: technical
relationshipSubType: metadata wrapper
relatedObjectIdentification: tif-138

The prototypes and experiments will drive the refinement of significant properties and relationships data model.

8. FUTURE DEVELOPMENTS

Even though many information units still require more investigation and the PML data model is still far from being finalized, the PML prototypes implemented in XML PREMIS semantic units will be published on the ARTAT website in late July 2010. Depending on the availability of the repositories technologists, tests will be performed on the understandability of the prototypes transmitted into their systems. In autumn 2010, the project will publish results about all workflows tested on the first participants. Hopefully, in the next year the ARTAT framework will be ready to welcome new participants.

Other developments on controlled vocabularies, MCO significant properties and relationships modelling will be integrated during the developing activities, as well as the feasibility of adopting semantic web technologies which could empower the shared preservation metadata for project's partners' advantage.

9. AKNOWLEDGEMENTS

The authors would like to thank Laura Ciancio of ICCU and Luca Lelli MAGTECA technologist, Giovanni Bergamin of National Library of Florence and Raffaele Messuti of Magazzini Digitali and Alessandra Giovenco of BSR who collaborated enthusiastically during the inquiry phase.

10. REFERENCES

[1] Caplan, P., Kehoe, W., Pawletko, J. (2009). "Towards Interoperable Preservation Repositories (TIPR)", *Proceedings of the iPRES 2009: the Sixth International Conference on Preservation of Digital Objects*, San Francisco, USA, 2009. Retrieved from: http://escholarship.org/uc/item/5wf5g5kh

[2] CCSDS, January 2002. *Reference Model for an Open Archival Information System (OAIS). CCSDS 650.0-B-1, Blue Book* (the full ISO standard). http://public.ccsds.org/publications/archive/650x0b1.pdf

[3] Di Iorio, A. (2009). "A Translation Layer to Convey Preservation Metadata", *Proceedings of the iPRES 2009: the Sixth International Conference on Preservation of Digital Objects*, San Francisco, USA, 2009. Retrieved from: http://escholarship.org/uc/item/4219t4n1

[4] Grace, S. (2009). "Investigating the Significant Properties of Electronic Content over Time final report", King's Inspect Project, JISC, College London, The National Archives, http://www.significantproperties.org.uk/inspect-finalreport.pdf

[5] Guenther, R., Wolfe, R. (2009). "Integrating Metadata Standards to Support Long-Term Preservation of Digital Assets: Developing Best Practices for Expressing Preservation Metadata in a Container Format", *Proceedings of the iPRES 2009: the Sixth International Conference on Preservation of Digital Objects*, San Francisco, USA, 2009. Retrieved from: http://escholarship.org/uc/item/0s38n5w4

[6] Jackson, A. (2006), "Preliminary Recommendations for Shareable Metadata Best Practices White Paper", Digital Collections and Content Project Grainger Engineering Library, University of Illinois at Urbana-Champaign. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.121.7095&rep=rep1&type=pdf

[7] Lagoze, C., Lynch, C., Daniel, R.Jr., (1996). "The Warwick Framework: A Container Architecture for Aggregating Sets of Metadata", Cornell Computer Science Technical Report TR96-1593. http://cs-tr.cs.cornell.edu:80/Dienst/UI/2.0/Describe/ncstrl.cornell%2fTR96-1593

[8] PREMIS Editorial Committee, 2008. *PREMIS Data Dictionary for Preservation Metadata version 2.0*. March 2008. http://www.loc.gov/standards/premis/v2/premis-2-0.pdf

[9] W3C Working Group Note (2006), "Defining N-ary Relations on the Semantic Web", www.w3.org/TR/swbp-n-aryRelations

Session 1b: Case Studies

ESA PLANS – A PATHFINDER FOR LONG TERM DATA PRESERVATION

Vincenzo Beruti **M. Eugenia Forcada** **Mirko Albani**
ESA-ESRIN
Via G. Galilei
CP 64,00044 Frascati
Italy
name.surname@esa.int

Esther Conway **David Giaretta**
STFC
Rutherford Appleton Laboratory
Didcot, Oxon
UK
name.surname@stfc.ac.uk

ABSTRACT

Digital preservation is difficult. The technical difficulties are the cause of much research. Other types of difficulty are those to do with organisational commitment, funding and context.

With the increasing interest on global change monitoring, also the use and exploitation of long time series of Earth Observation (EO) data has been increasing systematically, calling for a need to preserve the EO data without time constrains.

On the other hand:

- Data archiving and preservation strategies are still mostly limited to the satellite lifetime and few years after.
- The data volumes are increasing dramatically.
- Archiving and data access technology are evolving rapidly.
- EO data archiving strategies, if existing at all, are different for each EO mission, each operator or agency.

In the meantime the issue grows more urgent since more and more EO missions' data can be called 'historic' and more and more operators are faced with the decision of whether and how to preserve their data.

This paper describes the European Space Agency's (ESA) plans for long term commitment to preserving EO data concerning Europe. We believe this shows that ESA provides a pathfinder example of the way in which all these difficulties can be tackled.

1. INTRODUCTION

The need for accessing historical Earth Observation (EO) data series has significantly increased over the last ten years, mainly for long term science and environmental monitoring applications. This trend is likely to increase even more in the future in particular because of the growing interest on global change monitoring which is driving users to request time-series of data spanning 20 years and more, and due also to the need to support the United Nations Framework Convention on Climate Change (UNFCCC).

There are therefore strong drivers to preserve EO space data, keeping them accessible and exploitable for the long term. The preservation of EO space data can be also considered as a moral responsibility of the Space Agencies and other data owners as they constitute an asset for all mankind. In the next decade, the wealth of information currently locked inside the global data archives must be fully exploited and re-analyzed on a global scale (Figure 1 and 2 show examples of information extraction from long-term data series). This challenge relies on full accessible and exploitable archives (Figure 3).

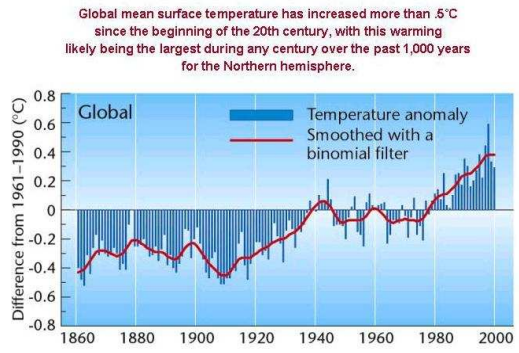


Figure 1. Surface Temperature increase

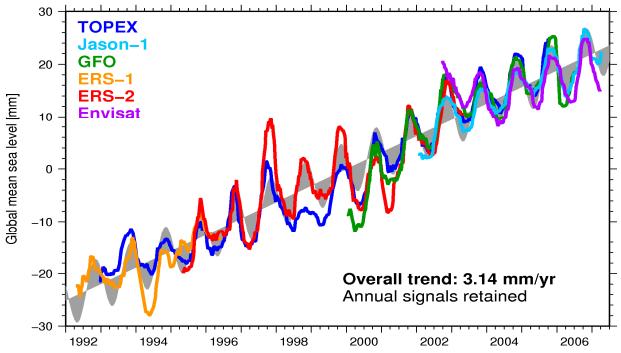


Figure 2. Global see level raise (courtesy of Remko Scharroo)

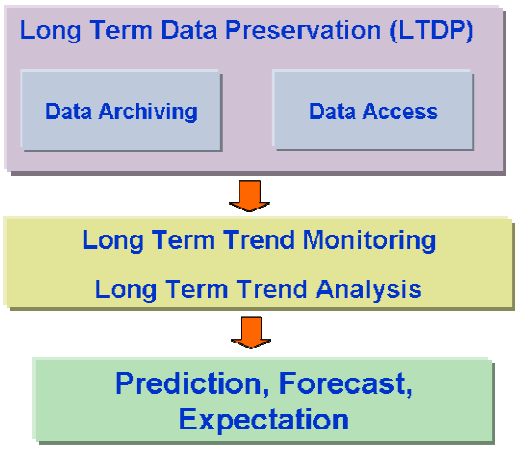


Figure 3. Long term trend monitoring

The application areas heavily benefiting from the EO long term data archiving exploitation are wide and can be summarized as:

- EC Policies with long-term perspective.
- European and Global Environment (e.g. Forest Monitoring, Soil Monitoring, Urban Development, Air Quality Monitoring, Ecosystems Monitoring and management for protection of terrestrial, coastal and marine resources).
- Management of energy resources (e.g. solar, etc.).
- Development and Humanitarian Aid Health including the understanding of environmental factors affecting human health and well-being.
- Food security including sustainable agriculture and combating desertification.
- Water resource management through better understanding of the water cycle.
- Civil Protection and disasters monitoring (e.g. Flood Prediction and Mitigation, Landslides, Subsidence, Volcanoes Monitoring).
- Global Climate Change (e.g. Systematic Climate Observations, Drought Monitoring, Monitoring of the Atmosphere, etc.).
- Climate understanding for assessing, predicting, mitigating and adapting to climate changes, as well as the improvement of weather information, forecasting and warning.
- Biodiversity enhanced understanding, monitoring and conserving.
- Global Security and Sustainable Development.

1.1. European EO archive challenges

The large number of new Earth Observation missions planned to come into operation in the next years (Figure 4) will lead to a major increase in the volume of EO space data. This fact, together with increased demands from the user community, marks a challenge for Earth Observation satellite operators, Space Agencies and EO space data providers regarding coherent data preservation and optimum availability and accessibility of the different data products.

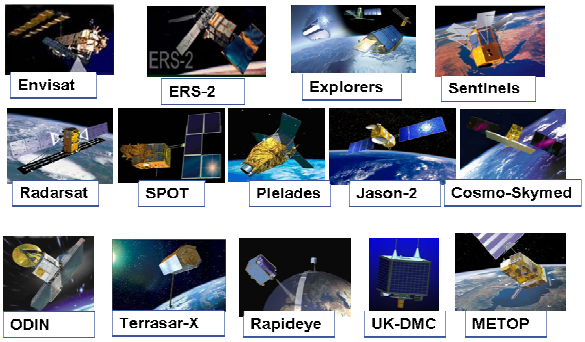


Figure 4. Current and future European and Canadian EO missions (excerpt)

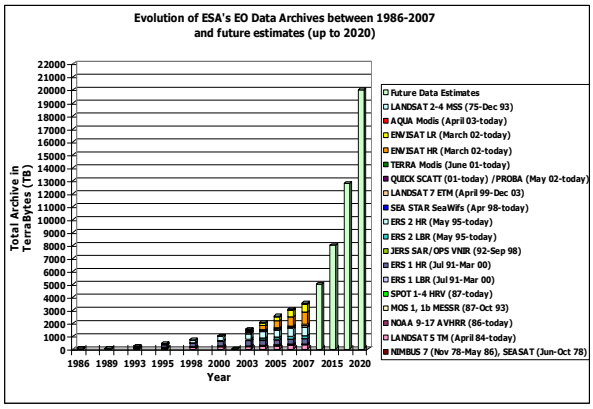


Figure 5. ESA expected volume of archives

As an example, Figure 5 shows the approximate current and predicted data volumes from a number of missions. The rate of growth of data volumes increased since the launch of ERS and in particular of the Envisat mission and will be increasing even further with the contribution of additional Third Party Missions (TPM) and with the GMES program. The plans of new ESA missions indicate 5-10 times more data to be archived in next 10-15 years. Similar trend is also monitored at all National archives.

Traditionally in Europe, there has been poor cooperation in this field with no common approach for long term preservation and access to EO space data despite the need for cooperation and sharing for the benefit of the user community. Preserving today's science records (e.g. data, publications) as well as their context is fundamental in order to preserve the future of science but single organizations have difficulties to afford data preservation in the long term that calls for the need of optimising costs and efforts, identifying commonalities.

A cooperative and harmonized collective approach on Long Term Data Preservation (LTDP) in Europe (i.e. a European EO LTDP Framework) is needed to coordinate and optimize European efforts in the LTDP field and to ultimately result in the preservation of the complete European EO space data set for the benefit of all European countries and users and with a reduction of overall costs.

2. ESA LTDP APPROACH

The European Space Agency now has a commitment to the long term preservation of its holdings and is putting in place the technical, financial and organisational wherewithal to accomplish this. Although at the time of writing not all the details have been decided, this paper describes some of the key principles which have been adopted. We believe that this approach can act as a pathfinder example for other disciplines.

In summary:

- A number of important principles have been identified and agreed with key Earth Observation space data holders and stakeholders in Europe and Canada.
- The key datasets to be preserved are being identified. This collection includes many types of digital objects, both data and documents.
- Common guidelines for key stakeholders have been defined in cooperation with European and Canadian EO space data stakeholders and are being refined.
- A technical and organisation framework is being prepared (European LTDP Framework).
- The way in which ESA LTDP fits into the broader international efforts in domains different from the Earth Observation one is being consolidated in order to maximise the usefulness of what ESA is doing and minimises duplication and waste of effort.

3. KEY PRINCIPLES

Main goals of the European EO Long Term Data Preservation Framework are to:

- Preserve the European, and Canadian, EO space data sets for an unlimited time-span.
- Ensure and facilitate the accessibility and usability of the preserved data sets respecting the individual entities' applicable data policies.
- Adopt a cooperative and harmonized collective approach among the data owners (LTDP Framework) based on the application of European LTDP Common Guidelines and sustained through cooperative (multi-source) long term funding schemes.
- Ensure, to the maximum extent possible, the coherency with the preservation of other non-space based environmental data and international policies.

The European LTDP Framework is open and is intended as a collaborative framework consisting of distributed and heterogeneous components and entities cooperating in several areas to reach a harmonized preservation of the European EO space data set. The framework is based on the contribution of European EO space data owners through their ideas and possibly their infrastructure in accordance to the commonly agreed LTDP Guidelines and should follow a progressive implementation based on a stepwise approach (short, mid, long-term activities).

A common approach in the field of Long Term Data Preservation should aim at the progressive application of the European LTDP Common Guidelines but also at cooperation of the archive owners in several areas for a progressive development and implementation of technology, methodology, standardization, operational

solutions and data exploitation methodologies as key aspects for the set-up of the framework. The European LTDP framework is outlined in Figure 6.

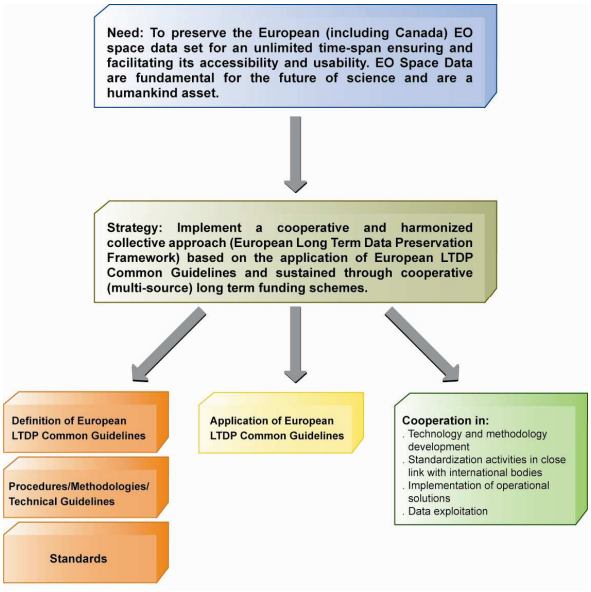


Figure 6. Long Term Data Preservation Outline Framework

A cooperative framework can facilitate for EO space data owners and archive holders the achievement of the common goal of preserving and guaranteeing access to the own data through benefiting from proven technologies, procedures and approaches and through the possibility to reuse and share infrastructure elements in the long term. The adoption of standards (e.g. for data access interfaces and formats, procedures, etc.) and common technical solutions can also allow to significantly reduce preservation costs.

The European LTDP Framework should be sustained through a cooperative programmatic and long term funding framework based on multilateral cooperation with multiple funding sources from at least the European EO space data owners.

The existence of a European LTDP Framework will also increase the awareness on data preservation issues favouring the start of internal processes at private or public European EO space data owners and providers. A European framework could also trigger the availability in the long term of additional permanent funding sources (e.g. European Commission) and can increase the possibility for any European (including Canada) EO space data owner to preserve missions data beyond their funding schemes into the cooperative and distributed framework.

4. THE LTDP COMMON GUIDELINES

In 2006, the European Space Agency (ESA) initiated a coordination action to share among all the European (and Canadian) stakeholders a common approach to the long term preservation of Earth Observation space data. During 2007, the Agency started consultations with its Member States presenting an EO Long Term Data Preservation strategy [3] targeting the preservation of all European (including Canada) EO space data for an unlimited time-span ensuring and facilitating their accessibility and usability through the implementation of a cooperative and harmonized collective approach among the EO space data owners.

The Long Term Data Preservation Working Group with representatives from ASI, CNES, CSA, DLR and ESA was formed at the end of 2007 within the Ground Segment Coordination Body (GSCB, [5]) with the goal to define and promote, with the involvement of all the European EO space data and archive owners, the LTDP Common Guidelines and also to increase awareness on LTDP. The resulting draft LTDP guidelines were reviewed by all ESA member states in the DOSTAG. The scope of Long Term Data Preservation as intended in the guidelines is not limited to the preservation of the data in the archives but also of the capabilities to generate products from the archived data and includes therefore also processing aspects. The insurance and facilitation of access, respecting the individual entities applicable data policies, and exploitation of the archived data are also part of the guidelines. Data access policies are on the other hand not part of the European LTDP Common Guidelines.

During the 1st Earth Observation Long Term Data Preservation workshop in May 2008 [6], the draft guidelines and the framework were presented and debated by all European and Canadian EO data owners, data providers and archive holders. The participants discussed and developed a joint strategy to move ahead technically and programmatically concerning the Long Term Data Preservation of EO Data and recognized the need and benefits of a common approach. Furthermore all the participants identified and agreed the draft LTDP Common Guidelines presented at the workshop as a first concrete and fundamental step to move ahead in creating the Long Term Data Preservation Framework. The guidelines should be adopted for old missions with a step-wise approach and straightforward for new missions and projects. ESA was given the task to trigger and coordinate the following steps toward the progressive European LTDP Framework implementation.

A consolidated LTDP Common Guidelines document has been produced on the basis of the comments and feed-backs received during the LTDP workshop. The document addresses the following nine main themes defining for each the “Guiding Principle” and the “Key Guidelines”:

- Preserved data set composition
- Archives maintenance and data integrity
- Archives operations
- Data security

- Data ingestion
- Data access and interoperability
- Data exploitation and re-processing
- Standardization
- Data Purging/Appraisal

An extensive public review process of the guidelines document was undertaken which collected additional comments and feed-back from the EO space data owners and archive holders. The review process has been completed and the guidelines document can be found at [3], but they are under continued review.

The LTDP guidelines constitute a basic reference for the long term preservation of EO space data. Their application by European EO space data owners and archive holders is fundamental in order to preserve the European EO space data set and to create a European LTDP Framework. The application of the identified guidelines is not mandatory for European EO space data owners and archive holders but is strongly recommended following a step-wise approach starting with a partial adherence. To this end different priorities have been associated to each guideline and three different levels of adherence to the LTDP Common Guidelines as a whole have been defined. Adherence to the guidelines should start from the basic level ones to reach full adherence in the long term. The LTDP guidelines document is intended to be a living document and can be also considered as a starting point to support the establishment, and aid the implementation, of more detailed technical procedures/methodologies when missing, favouring active cooperation in Europe in the LTDP field.

5. COOPERATIVE ACTIVITIES

The initial areas of cooperation related to LTDP to be addressed are on:

- Policies for the consolidation and issue of the European LTDP Common Guidelines, and adherence to them, and for the definition and application of a purge alert / appraisal procedure to EO space data.
- Technology, methodology and developments. The aim is to jointly evolve archive and data access technology through studies/pilots sharing the acquired know-how and infrastructure and to share knowledge/experience exchanging information to favour technical cooperation (cross participation into reviews, share of solutions, products, developments, etc). An additional fundamental activity is the continuation of development of harmonized access mechanisms (like HMA) and the definition of common operational procedures.
- Standardisation activities in close link with international bodies (e.g. CCSDS, CEOS, OGC, INSPIRE, EU initiatives, GEO).
- Operational solutions setting-up the principles for a common European distributed archiving concept aiming at the creation of an interoperable network of archive centres possibly reusing infrastructure of the different entities (as a single archive) in the long term. This would pave the way to future cooperation programs starting from standardized and certified services (e.g. share of data archives, archive transfer on demand or in case of a purge alert, coordination of re-processing schemes, format adoption/conversion, etc.).
- Data exploitation through the definition and implementation of joint EO historical data exploitation programmes.

An initial set of activities has been started in some of the areas mentioned above (e.g. studies on next generation archive technology and for the definition of LTDP users' requirements and composition of the data set to be archived to guarantee knowledge preservation). Details on the activities and their results will be provided published on the LTDP area of the GSCB web site [6]. Participation to the cooperation activities is open to European EO space data owners and archive holders and based on voluntary contribution.

6. TECHNICAL ROADMAP

ESA has developed the Standard Archive Format for Europe (SAFE) [10] an extension of the XFDU standard [11]. SAFE has been designed to act as a common format for archiving and conveying data within ESA Earth Observation archiving facilities. Many of the most important datasets have been converted to this format. The important point is that XFDU, and therefore SAFE, is designed to implement the OAIS Archival Information Package [9], which in principle has everything needed for long term preservation of a piece of digitally encoded information.

Some of the other components under consideration for the ESA LTDP technical implementation are the hardware needed to store the large volumes expected. Detailed studies of available and near-term storage solutions will inform the decisions which need to be made as a matter of urgency.

ESA has been involved with several EU part-funded projects concerned with digital preservation. Of these the PARSE.Insight project provides a roadmap [8] which is built on a broad survey of researchers, publishers and data managers [7].

The CASPAR project [1], in which ESA played an important role, has identified a number of key preservation components which should form part of a

shared infrastructure. These components will be evaluated as part of the ESA LTDP Framework.

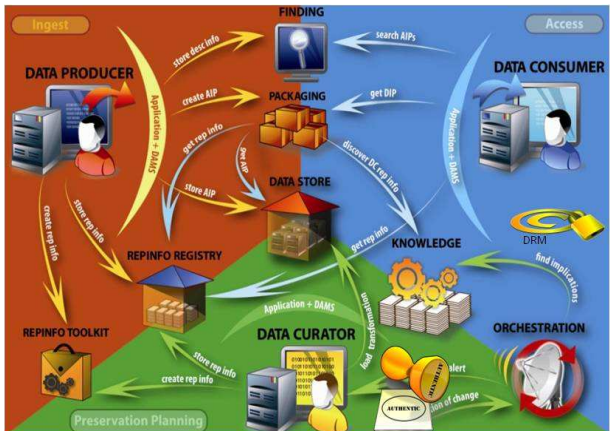


Figure 7. CASPAR workflows

For example when a data producer which to archive data, an AIP is created using the Packager. The AIP contains Representation Information, possibly from a Registry of Representation Information (RepInfo) or created with the RepInfo Toolkit. The amount of RepInfo depends upon the chosen Designated Community and the Knowledge Manager helps with this. Information supporting Digital Rights (DRM) and Authenticity are also needed The AIP is deposited in the Data Store.

More broadly one can view projects funded within the EU which form the basis of a Science Data Infrastructure (SDI) (taken from Mario Compalargo).

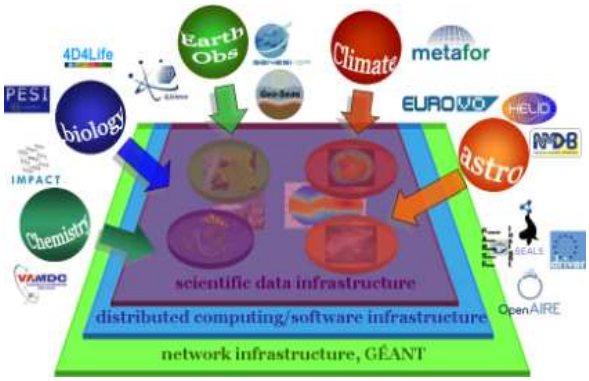


Figure 8. Landscape of e-infrastructure projects in the SDI area

Figure 8 illustrates the view that there should be a number of aggregators, including Earth Observation, Astronomy and Chemistry. These will be brought together into an overall science data infrastructure, which must include infrastructure components which assist in the preservation of digitally encoded information.

7. ORGANISATIONAL ROADMAP

The roadmap for the set-up of the European LTDP Framework can be articulated in three main phases to follow the best practices defined in the LTDP Common Guidelines and to progressively cooperate in the areas defined in Section 5. The initial situation is characterized by weak standardization, no clear and common methodology, poor cooperation with each entity dealing with the preservation of its own data often in a one by one mission basis.

The objective of the first phase is to reinforce the LTDP approach at each entity, to start the cooperation among agencies and EO space data owners (e.g. in methodology, standardization, sharing of information, etc…) and to define the future European organisation of LTDP with a very long term perspective. At the end of the first phase, standardization and methodology should be greatly defined on the main areas, EO space data owners should deal with LTDP as a transversal activity not closely tied to single missions and good cooperation among a significant number of EO space data owners should be in place.

In the mid term perspective (second phase) cooperation should be strengthened through the implementation of common activities among European EO space data owners with the goal to achieve an interoperable network of archives (e.g. share of solutions and systems, coordination of common technology developments, adoption of standards…), but also improving operational services according to user needs (i.e. Climate changes monitoring operational systems, etc.). Additional entities in Europe will be attracted and become part of the European LTDP Framework that is characterized at the end of the second phase by common technical views and solutions and standardized services, coordinated LTDP approaches and schedules between members and high interoperability between archives. LTDP is at this stage a common process which may be offered as a service able to manage and preserve the huge amount of EO space data owned by the different entities with a cost effective approach.

In the long term perspective cooperation should be further extended through the sharing of infrastructure (e.g. common and shared access points, interoperable and transparent data access and infrastructure), allowing to have a unique network of data, shared resources for data reprocessing and products generation and a common and harmonised security levels and layers. At this stage the EO space data archives could be extended also to other types of data according to user needs and alternative scenarios for the management of the exabyte/zetabyte archives era could be analyzed.

8. CONCLUSIONS

The European Framework for LTDP has been initiated. All European EO space data owners and archive holders are becoming part of a progressive process to guarantee the EO space data preservation in the long term. The European LTDP Common Guidelines have been issued and published on the GSCB web site and their promotion within CEOS or GEO international communities will be performed in the near future. The first identified technical cooperation activities have been started (e.g. studies on next generation archive technology and for the definition of LTDP users requirements and composition of the data set to be archived to guarantee knowledge preservation) and methodologies and standards available in international committees (e.g. CCSDS) or generated in the framework of ongoing projects (e.g. EC funded) are being revised for possible recommendation for adoption within the LTDP Common Guidelines. The international context of the European LTDP Framework is shown in Figure 7. In November 2008 at the ESA Ministerial Council, an LTDP programme for the period 2009-2011 was approved and ESA is now planning to apply the LTDP Common Guidelines to its own missions. ESA will implement the high priority activities in the next three years focussing on data preservation and enhancement of data access; an LTDP programme proposal will be prepared for the period beyond 2011.

9. REFERENCES

[1] CASPAR project web site http://www.casparpreserves.eu/

[2] European LTDP Common Guidelines, Draft Version 2, http://earth.esa.int/gscb/ltdp/EuropeanLTDPCommonGuidelines_DraftV2.pdf

[3] European Strategy for Long term EO data preservation and access, ESA/PB-EO/DOSTAG(2007)2, 8 October 2007.

[4] First LTDP Workshop, May 2008, http://earth.esa.int/gscb/ltdp/LTDP_Agenda.html

[5] GSCB Web Site, http://earth.esa.int/gscb

[6] LTDP area on GSCB Web Site, http://earth.esa.int/gscb/ltdp

[7] PARSE.Insight general survey report http://www.parse-insight.eu/downloads/PARSE-Insight_D3-4_SurveyReport_final_hq.pdf

[8] PARSE.Insight Roadmap http://www.parse-insight.eu/downloads/PARSE-Insight_D2-1_DraftRoadmap_v1-1_final.pdf

[9] Reference Model for an Open Archival System (ISO14721:2002), http://public.ccsds.org/publications/archive/650x0b1.pdf or later version. At the time of writing the revised version is available at http://public.ccsds.org/sites/cwe/rids/Lists/CCSDS%206500P11/Attachments/650x0p11.pdf or elsewhere on the CCSDS web site http://www.ccsds.org

[10] SAFE web site http://earth.esa.int/SAFE/

[11] XFDU standard available from http://public.ccsds.org/publications/archive/661x0b1.pdf

PRESERVATION OF DIGITISED BOOKS IN A LIBRARY CONTEXT

Eld Zierau **Claus Jensen**

The Royal Library of Denmark
Dep. of Digital Preservation
P.O.BOX 2149
1016 Copenhagen K, Denmark

ABSTRACT

The focus of this paper is on which digital objects to preserve when preserving digital library materials derived from original paper materials. It will investigate preservation strategies for digital objects from digitised paper material that must both be preserved and simultaneously retain a short route to dissemination. The investigation is based on a study of digitisation done a decade ago and digitisation done today.

In the last decade mass digitisation has become more commonly used since technological evolution has made it cheaper and quicker. The paper explores whether there are parts of digital material digitised a decade ago worth preserving, or whether a re-digitisation via mass digitisation today can create a relevant alternative.

The results presented show that the old digitised objects are worth preserving, although new digitisation can contribute additional information. A supplementary result is that investment in digitisation can mean lower costs in the long term. Manual adjustments for the image processing can result in considerably smaller images than images made in cheap mass digitisation. Although initial manual work is more expensive, the storage and bit preservation expenses are lower over a long period.

1. INTRODUCTION

This paper explores digital preservation of digitised material in a university and national library context, where there is a close relation between preservation and dissemination of digital material. The study is a result of a research project at the Royal Library of Denmark where the goal is updated preservation strategies for the library. It uses the Archive of Danish Literature (ADL) as a case study, which is a web based framework built at the start of the century. ADL is mostly limited to books, book metadata and book collections. In order to analyse the difference between 2000 and 2010, we have done experiments on re-digitising books from ADL.

The hypothesis investigated in this study is that we can reuse existing data from digitisations (10 years or older). If this hypothesis holds, it will also mean that it will be economically beneficial to preserve the old data in the sense of preserving the investment of the early digitisation. The results of exploring the hypothesis will influence preservation and dissemination strategies for the Royal Library of Denmark.

During the last decade, many mass digitisation projects have taken place all over the world. Examples of use of mass digitisation by Google can be found in [2]. Another example is Norway's National Library using Content Conversion Specialists (CCS)[1]. A decade ago, the available technology imposed limits on how automatic, fast and cheap a digitisation process could be. Today mass digitisation can be done much more cheaply and rapidly. However, there is no straight forward way to see if there is a difference in quality of the produced digital material. Quality according to requirements is important for whether digital material is worth preserving, therefore the differences will influence the strategies for preservation.

The study will explore preservation strategies mainly regarding functional (logical) preservation aspects of digitised objects, where a digital object must be preserved to be understandable and usable in the future. But functional preservation related to representation of complexities like consecutive pages is not part of this paper. The underlying bit preservation, which must ensure that the actual bits remain intact and accessible at all times, is only mentioned briefly.

Dissemination must be taken into account when evaluating preservation options in a library context. In comparison with e.g. traditional archives, libraries face additional challenges to preservation, since digital material in many cases must be disseminated to the public or researchers through fast access. Differences in purposes and goals for dissemination and preservation place different demands on the formats in which digital materials are preserved and presented, respectively. For example, many libraries have chosen TIFF or JPEG2000 as the preservation format for books and images [3], [8].

[1] See http://newsroom.ccs-digital.info/index.php?option=com_content&task=view&id=12&Itemid=26

Dissemination, on the other hand, may use formats that consume less storage, e.g. JPEG or GIF, or formats with additional information for dissemination, e.g. pyramid-TIFF[2] (derived from TIFF) or JPEG2000 for images needing zoom functionality.

Besides the influence of dissemination requirements, the study will evaluate the choice of digital objects for preservation on different parameters. These are; quality of contents of the digitised object, the ability of the format to be used as a preservation format, the cost of producing objects, size of objects (related to ongoing storage costs), and the risks associated with the choices for production and storage of the digitised objects.

2. CASE STUDY: THE ADL SYSTEM

In order to explore the hypothesis, we will use the Danish ADL System as a case study. This system was developed by the Royal Library of Denmark together with "Det Danske Sprog- og Litteraturselskab" (DSL) which publishes and documents Danish language and literature. The Royal Library developed the framework, while DSL selected literary works to be included. The ADL system is a web based dissemination framework for digitised material from the archive for Danish literature. Today it contains literature from 78 authors represented by over 10,000 works of literature (e.g. novels, poems, plays). ADL additionally contains author portraits as well as 33 pieces of music (sheet music) and 118 manuscripts. The publication framework is still available on http://www.adl.dk/.

The case study is interesting because it reflects a system built on the basis of technologies from the start of this century. Today, new demands have arisen for dissemination and preservation, and new technologies exist to produce the digital material.

2.1. Present Architecture and Contents

The ADL system does presently offer display of book pages based on the framework designed a decade ago. Each page can be viewed in three different ways derived from original scanned TIFF files with page images; as a 4-bit GIF image, as a pure text representation or as a download of a PDF containing the page image for print. This is a typical application from 2000 where 4-bit GIF was chosen to allow quick dissemination of ADL web-pages to users using relatively slow connections.

Digitised manuscripts and sheet music were added at a later stage. These are represented via JPEG images, because JPEG is a better dissemination format for e.g. handwriting on yellowed/coloured, deteriorated paper.

The structure of the web framework is based on authors, their literary works and the period when the authors were active. The website is based on dynamic HTML pages generated from information in a database.

2.2. Present versus Desired Preservation

The focus is on the preservation of the digital objects, thus we will not go into details of the functionality of the system. It should however be mentioned that scalability for fast response time and ease of maintenance of dissemination applications must be taken into account in the final decision on the preservation strategies.

At present, the ADL is only preserved as a part of the Danish web archive. This means that the only data preserved is the data visible on the internet, which does not include e.g. TIFF files and special encodings. Further actions for preservation await the results of this research project.

The preservation strategy considered for ADL data is a migration strategy. The focus here is to preserve and possibly reuse earlier digitisation as a basis for a migration into emerging dissemination and preservation formats. Emulation[3] does not support changes in presentation form and is therefore not considered.

The ADL information which is the target for preservation is: the digitised representation of the book *items* (as defined in [10]) including page images and encoded texts, manuscripts and sheet music as well as the related information such as period descriptions and author descriptions. Since the author and period descriptions were written especially for the ADL system, these are born digital.

ADL is presently disseminated from its own platform which is not aimed at preservation. This will change when a preservation strategy is implemented for ADL. Dissemination in a library context is strongly related to preservation. For example, if in dissemination we use high consumption storage formats similar to the preservation format, we may want the preservation and dissemination modules to share a copy of the data, or to be able to produce dissemination copies quickly for a cached storage. Sharing a copy under bit preservation should however be done with care (see [13]). Deriving a dissemination copy requires that it will be possible to identify and retrieve preserved data on request. Furthermore, the cost of transforming data for dissemination must be minimal. In the long term, a shift from e.g. TIFF to JPEG2000 in dissemination must be coordinated with the preservation formats, and vice versa, to support scalability and efficiency.

3. EXPERIMENT SETUP

The experiments focus on the issues related to the digitisation of book items. This excludes manuscripts and sheet music, author descriptions, period descriptions, relations between information such as citations etc. The book items included have good print quality. In ADL only the text and how the text is

[2] TIFF, Pyramid. The Library of Congress (National Digital Information Infrastructure & Preservation Program)

[3] See e.g. "Keeping Emulation Environments Portable" (KEEP). http://www.keep-project.eu/

expressed through layout and text structure are worthy of preservation. This means we do not view look & feel and illustrations as important.

The goal of the experiments is to investigate how the book information should be preserved, when we consider the costs, available technology and the higher demands for dissemination. The experiments focus on questions related to our hypothesis that the quality of the original material is such that it provides a solid basis for future development.

3.1. Preservation Scenarios – Data to be Preserved

We will investigate our hypothesis in terms of the value of the original data from ADL compared to the value we can get from a re-digitisation. On this basis we can explore different preservation scenarios that can be used in a preservation strategy for the library.

The data we will investigate is:

Book item, which must be preserved, if a later rescan can be expected to add value in form of extraction of additional information or substitution.

OCR and encoded text from the original digitisation must be preserved, if it contains information that is expensive or hard to recreate.

Low resolution page images from the original digitisation must be preserved in case we conclude that we do need page images, but not necessarily in high resolution.

High resolution page images from the original digitisation must be preserved in case we expect to do future OCR adding new information, or in case we expect to do manual inspection on letters that are hard to read. Look & feel and illustrations for dissemination can also be issues, but not in the ADL case.

The preservation scenarios considered are defined in terms of combinations of data we choose for preservation. The scenarios are listed in Table 1.

	Scenario						
Data	1	2	3	4	5	6	7
Book item	*X*				*X*		
High res. page image		*X*				*X*	
Low res. page image			*X*				*X*
OCR & encoded text				*X*	*X*	*X*	*X*

Table 1. Scenarios 1-7 for preservation of data.

For the column representing scenario 1, Table 1 has an 'X' in the row with book item. This means that in scenario 1, only the book items will be preserved, thus the digital preservation is skipped. Likewise, Table 1 shows that in scenario 6 both OCR & encoded text and high resolution page images are preserved.

Choice of scenario will be evaluated against what data we need to preserve, the associated risks, the expected costs, and consequences of choice of preservation format with respect to e.g. maturity and available tools.

3.2. Digitisation Process

To better understand the set-up of the experiments, we here give a brief sketch of the digitisation process as it was performed with the original ADL data. The process, illustrated in Figure 1, was defined on the basis of the then current experiences and observations made in e.g. the DIgitised European PERiodicals (DIEPER) project [9] and best practices within digital imaging [7].

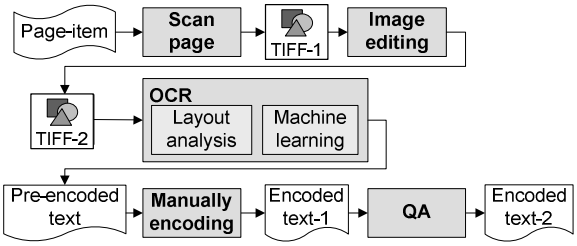


Figure 1. Simplified digitisation process.

Not all the sub-results illustrated in Figure 1 are available in the ADL case, since some data has been deleted or modified. The available information is; the page items, the image edited TIFFs (TIFF-2) and the encoded text after quality assurance (Encoded text-2).

Page item(s) from books. In ADL the books had their backs cut off, because the scanners used were document scanners with automatic page feeding.

Scan page(s) of page items. In ADL scanning was optimised and adjusted to get the best quality of the pages [9]. Grey-tone scale or adjusting the depth was used. Most pages in ADL were scanned with TIFF - 400 DPI Grey scale, 400 DPI black/white, a few with 600 DPI- Black/white and some with 200-300 DPI for pages with a high degree of background noise.

Image editing to deskew, image centering, and light & contrast adjustment. In ADL these edits were made to get a better presentation in dissemination. Furthermore, deskewing and adjustment of light and contrast enhanced the OCR results.

OCR, Optical Character Recognitions. In ADL the OCR was performed with FineReader version 5.0 or 6.0.

OCR - Layout analysis of pages. In ADL this analysis was used to identify and record objects types like text blocks, images, tables etc. and their appearance order. Common errors for the ADL scans were that the objects were not identified or they were identified with wrong type, e.g. as an image instead of a text. Manual analysis and predefined blocks for objects were used as help.

OCR - Machine learning of letter recognition. In ADL this covered assignment of Danish dictionaries that the OCR was mapped against.

Pre-encoded text results from the OCR. In ADL this included automatic formatting with page-breaks, line-breaks, new paragraph, etc.

Manually encoding as additional manual encoding result to the OCR result. In the ADL case, the pre-encoded text, represented in XML files, was sent to

Russia, Sweden or India for manual encoding of various TEI-P4/TEI-lite[4] codes including e.g. speakers in plays.

QA, quality assurance of results. For ADL data this was mainly made automatically by the upload script to the ADL database which carries out a syntax check of the XML. On this point the best practices had not been followed.

3.3. Experiment Selection

Six books were chosen for the re-digitisation experiment based on wide representation of the following criteria:

- Aspects in recognition of characters. For example black-letter[5] typeface is hard to recognise for the OCR, and the results may differ for different fonts and print types
- The genres of the books. For example plays are harder to encode than novels and poems
- Illustrations included in the book. These can present challenges to the OCR block recognition
- Notes in footer or margin in the book. These can present challenges for text encoding of the notes

The books chosen were:

{a} *Ludvig Holberg, Værker Bd. 1*. It contains essays and marginal notes. Marginal notes are not included in the current ADL encoded text.

{b} *Ludvig Holberg, Værker Bd. 3*. It contains plays and images. The book item was exchanged with another copy of the same edition. The only difference is a stamp on page 4 in the original.

{c} *Henrik Hertz, Dramatiske Værker Bd. 3*. It contains plays and is printed in a black-letter typeface.

{d} *Karl Larsen, Doctor Ix.* It contains a novel, and the print seems a bit thin.

{e} *J. P. Jacobsen, Lyrik og Prosa.* It contains poetry and diaries. Pages 96 and 97 are missing in ADL.

{f} *Anders Bording, Samlede Skrifter DDM*. It is printed in a black-letter typeface, and was included only because OCR had originally been given up.

The re-digitisation was carried out in two places and carried out with two different approaches. One represents cheap mass digitisation and the other represents re-digitisation as done for the original ADL data, and costing 8 times as much:

Subcontractor 1 (SC1) who carried out the mass digitisation. Here scanning was done in 400 DPI 24 bit colour set up in a standard configuration. OCR and text encoding was made in a semi-automated production via Germany-CCS docWORKS version 6.2-1.16 using FineReader version 7.1 and using ABBYY Morphology Engine 4.0 for an ABBYY dictionary. The results of the scanning were given in JPEG2000 and TIFF. The production of encoded text was based on TIFF and given in the ALTO[6] xml format. The automatic process did not involve image editing, manual encoding, special setup of OCR layout analysis and special setup of machine learning (see Figure 1).

Subcontractor 2 (SC2) who used an approach similar to the original ADL digitisation. OCR and text encoding based on the original ADL TIFF files, i.e. without scanning of page items and Image editing (see Figure 1). SC2 made OCR by FineReader version 9 and encoding in TEI-P4 with manual codes as specified for the original encoding. The process started on the basis of 'TIFF-2' from ADL (see Figure 1).

We decided only to do experiments on scans of the books via SC1. The interesting part is to see what the differences are between the original scans and scans made in an automatic process with standard scanning configuration for all books. We did not expect to see many differences because of the good quality of the ADL material, which is also the reason why scans were not part of the SC2 setup. The OCR set-up will in most cases still work best on scans with grey-tone 400 DPI [9], [11]. In the SC1 case the scans are in colours. Despite this, the cheap digitization price and the assumption of small differences made us settle for SC1.

Book {f} was only sent to SC2. The reason for this was that the encoding of this book had been given up earlier, thus it could not be expected to give a better result in an automated process. The most interesting investigation in this case is to see if improved technology enables OCR and text encoding of {f} today.

Encoding of marginal notes excluded in ADL from book {a} was included both in the SC1 and the SC2 results. This forms a basis for investigating if the new encoded files can replace the original encoded ADL files, because of added value of the margin encodings.

We did additional in-house experiments with JPEGs in order to investigate the question whether images can be preserved in a format requiring less storage space with less quality. For this experiment we made a sample selection of pages from the original ADL TIFFs and JPEGs derived from these TIFFs. Corresponding pages from the JPEG2000s and JPEGs received from SC1 were used. These experiments were conducted using FineReader version 10.

4. EXPERIMENT RESULTS

In this section we will start by presenting the experimental results compared to the original ADL data. Next we will relate these results to the different preservation scenarios given in Table 1.

4.1. Scanning Results

The new scans have an acceptable quality, but differ in adjustment, number of pages, colours and storage sizes. The quality is acceptable in the perspective that letters are readable from a screen presentation, i.e. it can be used for dissemination and proofreading. Furthermore, as described later, the scans can be used as basis for

[4] TEI (Text Encoding Initiative)
[5] See http://en.wikipedia.org/wiki/Blackletter
[6] ALTO (Analyzed Layout and Text Object). 2004. Technical Metadata for Optical Character Recognition, version 1.2.

OCR. Illustrations are not important here, but it can be noticed that in dissemination used for ADL, they appear similarly to the old ADL scans.

The adjustment difference is due to lack of an image editing process in the automatic scans (see Figure 1). The same reason applies to the extra pages in the SC1 scans, since blank pages or pages with edition information have been removed in the original ADL image editing process.

The reason for the difference in colours is that the new scans are done in colour, while the originals were made in grey tone or black and white. This is also part of the explanation for difference in the storage sizes. However, the increased storage size is also caused by extra margins (thus larger image) which are removed in the editing process of the ADL scans, and the extra depth in some of the SC1 scans compared to ADL scans.

Table 2 gives an overview of the difference in sizes compared to the ADL scans. Note that there can be variations in these numbers depending on character density, original ADL scanning technique etc.

Format	Storage factor of ADL TIFFs
SC1 TIFF	*10 times bigger*
SC1 JPEG2000	*2 times bigger*

Table 2. Storage space factor of page format.

Besides size difference for the individual pages, the SC1 will require extra storage space for the extra pages which were deleted in the ADL editing process, and the missing pages for book {e}.

4.2. OCR Results

The detailed OCR results are based on samples of pages selected on basis of variations in the page layouts.

4.2.1. .Latin Typeface Pages in OCR from TIFFs

For books with Latin typefaces, the character recognition is fairly good as shown in Table 3. The numbers in Table 3 are number of differences (errors) in per mille where spaces and line breaks are excluded. For SC1 and SC2 the numbers are given for the errors in the OCR. For ADL the numbers are for the errors in the OCR with subsequent corrections.

Book \ Origin	ADL	SC1	SC2
{a}	0,1	1,4	1,2
{b}	0,0	2,5	1,3
{d}	0,0	3,3	2,0
{e}	0,3	7,5	3,7

Table 3. Number of errors (per mille).

The ADL OCR seems best, but has had subsequent corrections. Generally the difference between ADL, SC1 and SC2 OCR is small, therefore we cannot conclude whether one result is better than the other.

4.2.2. Latin Typeface Pages in OCR from JPEGs

The internal experiment with JPEG shows that the JPEG OCR was relatively good for Latin typefaces, as shown in Table 4 (calculated in the same way as for Table 3).

Origin	SC1		ADL	
Book	JPEG2000	JPEG	TIFF	JPEG
{a}	0,6	0,9	0,2	0,6
{b}	0,9	0,9	1,1	1,7
{d}	1,4	4,2	2,5	4,2
{e}	1,9	3,7	2,1	2,5

Table 4. Number of errors in the OCR (per mille).

Experiments with TIFFs from SC1 were also performed but the results were exactly the same as the results from the experiments with the JPEG2000s.

Most of the JPEGs have errors in letter recognition. Especially book {a} has many errors in the SC1 JPEG. In many cases the Danish 'ø' that is recognised as 'o'. We chose to study this problem further, an arbitrary 'ø' from the SC1 book {a} results, illustrated in Figure 2.

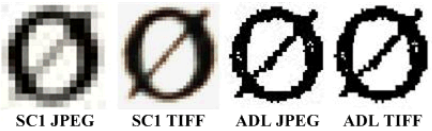


Figure 2. OCR of ø.

The images show that the ADL scans in TIFF are much sharper than the SC1 scans. One reason is due to optimisation in the ADL scanning and image editing (see Figure 1). In the conversion to JPEG the line in the 'ø' fades in the SC1 JPEG. This is not the same in the ADL JPEGs because they originate from TIFFs which are optimised in light and contrast.

The result gives an indication that a new encoding can be based on JPEGs for some books, although it will require extra quality assurance and manual corrections. The result also indicates that manual inspection can be based on the JPEGs for later corrections in the old OCR.

4.2.3 Black-Letter Pages in OCR from TIFFs

As expected the OCR of black-letter typefaces was not without problems. Some black-letter letters can be quite hard to distinguish even for a trained human eye. Examples are "d" and "v" as well as "f" and "s".

The OCR of black-letter text requires a special additional OCR program which was not part of the SC1 package, therefore the OCR results for these books are not interesting. The SC2 faced challenges with both books which differ in black-letter fonts, and in print quality. It required addition of a special Danish dictionary and manual work to get a reasonable result.

OCR of book {f} had been given up earlier, and it did give SC2 additional challenges. Especially black-letter capitals were hard to recognise in this book. The result includes about 15-20% errors in character recognition, and many black-letter characters are interpreted as

images. The earlier attempt resulted in approximately 40% errors in character recognition. Although the new OCR results are an improvement, there is still a need for a lot of manual work in order for the text to be acceptable for dissemination.

Book {c} had been through OCR and text encoding earlier with success, and SC2 has produced a much better result with less than 1% errors. Still the ADL is better with only few errors. This does not lead to conclusion that the ADL OCR was better, since it can be cause by subsequent corrective actions in the ADL XML. The comparison was made by manual inspection in order to determine whether the SC2 and/or ADL text is wrong. As described in "Improving OCR Accuracy for Classical Critical Editions" [1] there is no way to determine this automatically.

4.3. Encoding Results

The encoding results are in different formats. The SC1 result is given per page in ALTO which is an XML representation of automatic derivable typographical information about the appearance of words in the layout, e.g. paragraphs, line breaks, word positions etc. The SC2 is given in TEI-P4 which additionally contains text structural information such as interpretations of a chapter, a paragraph, a line group, a poem, a literary work, a speaker etc. However the SC2 TEI-P4 has no position information and it follows a different kind of XML tree structure than ALTO.

The different file formats and contents also influence how much storage space the files require. Due to the very detailed positions information in the SC1 files, these files require about 20 times more storage space than the ADL files. This will, however, vary according to text density on pages and inclusion of illustrations.

4.3.1. Typographical Encodings

It was not straightforward to compare the SC1 ALTO files with the SC2 TEI encoded files, because the representations are so different. However, we found that most of the information in the ALTO files is included in the TEI files (line breaks, paragraphs etc.).

Positions cannot be compared except for accuracy, even if there had been positions in the SC2 results. The reason is that positions from SC1 are related to the scans in the SC1 result and thus are very different from the original ADL scans. This means that the ALTO results only are valuable if SC1 scans are preserved as well.

The results from book {a} with marginal notes are noteworthy. These notes were left out of the original digitisation, and can therefore only be compared between the SC1 and SC2 results. In the results from SC1 the marginal text was encoded separately from the section text, and marked as marginal text with the specification of the position of text blocks and individual words. In the results from SC2, marginal text was placed above the text-section that the note belonged to. This is not very precise since the notes have a more specific placement in the layout.

4.3.2. Text Structural Encodings

Generally, the results from SC2 do not have as good a quality as the original ADL encodings. For instance the stage directions in drama, introducing and ending a scene, are encoded in the ADL text, but only given as italic encoding of each line in the SC2 text. Hyphenation is encoded in several of the ADL texts, but none in the SC2 text, the same applies for line groups.

There are big differences in the use of TEI-codes between the TEI-files in ADL and the corresponding TEI-files from SC2. This difference will complicate merging the two results. For example, TEI div-tags were used in both files, but not in the same way and with different naming of the div tags.

5. GENERAL SCENARIO RELATED RESULTS

We will here look at the feasibility of the different scenarios. This will be done by evaluating the different materials used in the scenarios based on the results given in the previous section.

5.1. Book Items

For ADL, the only case where preservation of the book items is a necessity is when pages are missing. Here re-scans can create the missing pages. However, this also points at the importance of more thorough quality assurance of the scanned pages, which could eliminate the need to preserve the book in the ADL case. However for other kinds of material there will be cases where a re-scan is needed for other purposes. For example, if in the future higher resolution of images or look and feel of the page is needed. Thus a decision not to preserve the book items will add a risk of losing such information.

Reproduction of lost material from a digitised book will have to be based on the book item. This can be an expensive and time consuming process, especially if large chunks of data are lost, thus preserving book items alone will not meet the requirement for fast dissemination.

In the specific ADL case, the recommendation will be to ensure better quality assurance and possibly skip preservation of books. This rules out scenario 1 and 5. Note that this will need to be accompanied with a higher level of bit preservation. Furthermore, there will be many cases for digitised books in general, where it should be supplemented with preservation of the books.

5.2. OCR & Encoded Text

We can conclude that OCR & encoded text needs to be preserved. From the experiments we found that there is valuable information in the encodings which will be hard and expensive to reproduce in a re-digitisation process (especially text structural information, e.g. stage

directions and poetry line groups). Furthermore, loss of OCR & encoded text information will entail a long route to re-dissemination. Thus scenario 1, 2 and 3 are not to be recommended, since they do not include preservation of OCR & encoded text.

On the other hand, OCR and encoded text should not be the only information preserved. The reason is that there is too high a risk that information is wrong (spelling errors, fonts, italics, bold etc.) and cannot be detected or corrected by inspection of the book/page images. Another risk is that valuable information is lost (e.g. marginal notes) or because of inaccurate encoding (e.g. marginal note positions). Furthermore, it eliminates the possibility of asking the public for help to identify mark-up errors, as for example done in Australia [4]. Thus, scenario 4 cannot be recommended.

One should carefully consider how to preserve the OCR and encoded text. It needs to be preserved in a way that allows enrichment of the encoding and respecting how dissemination information can be derived. The final recommendation on how to preserve OCR and encoded text will therefore be closely related to consideration of modelling the book objects for logical preservation and preparing for future enhancements with annotations from the public and researchers. This work is described separately in [12].

5.3. Page Images

As long as the page images are needed as part of the dissemination, page images should be preserved, since loss of pages will mean a lengthy re-dissemination process or complete loss of pages. Furthermore, as described under book items and OCR & encoded text, page images are the best choice as a complement to preservation of OCR & encoded text. Thus the choice is between scenarios 6 and 7. The questions that remain is what image formats we can accept as a preservation format, at what cost, at what risk, and how to retain the possibility of a short route to dissemination.

A *preservation format* will need to be a well documented format, preferably loss-less, and supported by tools. If we look at a format like JPEG, this is a lossy format which loses data when edited. However, it may still be considered as a preservation format of page images, or perhaps loss-less format as e.g. GIF or loss-less JPEG[7] might be considered. As for TIFF, this is a simple, mature, high resolution format supported by many tools, but it consumes much storage space. JPEG2000 is a high resolution format, which requires less storage space, but is more complex, and not as mature and well supported as TIFF. For supplementary considerations see [3,8].

Costs can be related to the creation of the digital objects or to ongoing storage and maintenance in connection with bit preservation (e.g. cost of hardware migration and integrity checks between data copies [13]). Looking at the ongoing costs, one way to reduce the costs is to choose a format which requires less storage space; another way is to store the format in a compressed form. Table 5 gives approximate percentages of storage reduction compared to the ADL TIFF, including numbers for LZW[8] compressed TIFFs. Note that it does not make sense to LZW compress JPEGs, and that LZW works best on black & white.

Format	% Storage of ADL TIFF
ADL TIFF LZW	*8%*
ADL JPEG	*50%*
ADL XML	*0,1%*
SC1 TIFF	*(10 times bigger) 1000%*
SC1 TIFF LZW	*(2 times bigger) 200%*
SC1 JPEG2000	*(2 times bigger) 200%*
SC1 JPEG	*15%*
SC1 XML	*2%*

Table 5. Storage space factor of page format.

The percentages are based on the experiments, but will vary for each book because of differences in letter density and inclusion of images. Table 5 includes factors for LZW compression of the different formats.

The least storage consuming format is the LZW compressed TIFFs, which were created by an optimised scanning process. Since the SC1 JPEG2000 is twice as big as the ADL TIFFs, we can conclude that have gained considerable storage reductions through optimisation of the scanning process. However, compressed JPEGs may give a better result. To draw a general conclusion of whether an optimised scanning process will be cheaper, it must be investigated further if the extra costs for manual work in the optimised scanning process can compete with the cost savings in storage.

The choice of format must be evaluated against the risk that the format may add. For instance, if a format is a lossy format, there will be a risk related to whether transformation or edits have a negative effect. If page images are saved in low resolution, the risk is that it is too low for future needs, and for later automatic extraction of additional information. Furthermore, if all books are treated identically there is a risk that e.g. books printed with black-letter may have a higher risk of losing information, even using manual inspection. On the other hand, differentiated preservation strategies for different books will influence the complexity of preservation strategies. If we use compression of the formats, this will add a risk to the bit preservation. Furthermore, compression may also add processing time in dissemination.

For ADL, we end up with a recommendation of preserving LZW compressed TIFFs, since investment, in an optimised scanning process, has already been made, it is a stable format, and book items still exist, if compression corrupts the TIFF.

[7] See http://en.wikipedia.org/wiki/Lossless_JPEG

[8] See http://en.wikipedia.org/wiki/Lempel-Ziv-Welch

6. DISCUSSION

For this study we are privileged to have unique material from an application built a decade ago. However, the state of the books and lack of information about the original digitisation process has influenced how much we can conclude. It could be argued that the condition of the material added too many uncertainties to the results. However, no matter how the ADL data has achieved its good quality, we have been able to conclude that these data is worth preserving, and we have been able to analyse which preservation strategies to choose, thus the case study has fulfilled its purpose in the investigation of our hypothesis.

In this study we have only investigated digitised books with high print quality and only focused on the text, layout, and structure of the text. Other books with other characteristics may need other digitisation and preservation strategies. For example the page images of sheet music and manuscripts in colours may need higher resolution in preservation, and might give better results in mass digitisation. Another example is, if text structural information is unimportant then all information can be extracted automatically. In any case compromise and choices must be made at the start of scanning.

In the ADL case we saw that it seems beneficial, regarding ongoing storage costs, to digitise books using a more manual approach than mass digitisation traditionally takes. The actual difference in cost can be hard to evaluate. A model for calculation of migration costs is given in [5], and results will differ according to the period it covers and how it is used.

The JPEG case study only represents a special setup with specific parameters for OCR and conversion. The result may have differed with different tools and setups. However, similar results may be found in experiments with decrease of image quality.

Mass digitisation does have different advantages. For example, there may also be a time factor for how long a digitisation process must take, e.g. because of a political deadline or deteriorating material (an example can be found in [6]).

The preservation strategy must take into account how to model the complex structures so that the necessity of the short route to dissemination is respected. The reason is that it can influence how the metadata and the OCR and encodings are represented in the preserved data. In other words, the research presented here only provides input on the reuse of earlier digitisation, and which factors should be considered in new digitisation.

7. CONCLUSION

We can conclude that digital material from the ADL case study is worth preserving. In general it points at possible reuse of digitised books where look & feel as well as images are unimportant. The digital objects, that are the target for digital preservations are the page images and the OCR & encoded text. Whether the pages are preserved in high or low resolution, with or without compression, must depend on a risk analysis, and analysis of relation to dissemination.

An additional result of the study was that the chosen digitisation process can influence the ongoing storage costs in the future. This leads to a conclusion that the digitisation process must be chosen with care both regarding the immediate requirements, but also regarding the long term consequences.

Digitisation has evolved in the last decade, enabling cheaper and faster mass digitisation, as illustrated in the different digitisation approaches. The most evident evolution in the ADL case was enabling OCR of books printed in black-letter which were given up a decade ago. However, the manual work in scanning and in OCR & encoding corrections have added value to the material. For book material of as good quality as in ADL, the technological enhancements cannot compete with these manually added values.

A final preservation strategy in a library context can now be made on basis of this study as well as the modelling aspects related to functional preservation.

8. REFERENCES

[1] Boschetti, F., Romanello, M., Babeu, A., Bamman, D., Crane, G. "Improving OCR Accuracy for Classical Critical Editions" *Proceedings of the European Conference on Digital Libraries*, Corfu, Greece, 2009.

[2] Coyle, K. "Mass Digitization of Books". *The Journal of Academic Librarianship, Vol. 32, Issue 6*, 2006.

[3] Gillesse, R., Rog, J., Verheusen, A. "Life Beyond Uncompressed TIFF: Alternative File Formats for Storage of Master Image Files" *Proceedings of the IS&T Archiving Conference*, Bern, Switzerland, 2008.

[4] Holley, R. "Many Hands Make Light Work: Public Collaborative OCR Text Correction in Australian Historic Newspapers". *Technical Report from National Library of Australia*, Australia, 2009.

[5] Kejser, U.B., Nielsen, A.B., Thirifays, A. "Cost Model for Digital Curation: Cost of Digital Migration" *Proceedings of the International Conference on Preservation of Digital Objects*, San Francisco, USA, 2009.

[6] Kejser, U.B. "Preservation copying of endangered historic negative collections" *Proceedings of the IS&T Archiving Conference*, Bern, Switzerland, 2008.

[7] Kenney, A. R., Rieger, O. Y. *Moving Theory into Practice: Digital Imaging for Libraries and*

Archives. Research Libraries Group, California, USA, 2000.

[8] Kulovits H., Rauber A., Kugler A., Brantl M., Beinert T., Schoger A. "From TIFF to JPEG 2000? Preservation Planning at the Bavarian State Library Using a Collection of Digitized 16th Century Printings", *D-Lib Magazine Vol. 15 No. 11/12*, 2009.

[9] Mehrabi, H., Laursen H. "Standards for images and full text" *Proceedings of Conference on future strategies for European libraries*, Copenhagen, Denmark, 2000.

[10] Riva, P. "Functional requirements for bibliographic records: Introducing the Functional Requirements for Bibliographic Records and related IFLA developments" *Bulletin of the American Society for Information Science and Technology vol. 33 issue 6*, 2008.

[11] Holley, R. "How Good Can It Get? Analysing and Improving OCR Accuracy in Large Scale Historic

[12] Newspaper Digitisation Programs", *D-Lib Magazine Vol. 15 No. 3/4*, 2009.

[13] Zierau, E., "Representation of Digital Material Preserved in a Library Context". *Proceedings of the International Conference on Preservation of Digital Objects*, Vienna, Austria, 2010.

[14] Zierau, E., Kejser, U.B. "Cross Institutional Cooperation on a Shared Bit Repository". *Proceedings of the International Conference on Digital Libraries*, New Delhi, India, 2010.

RESHAPING THE REPOSITORY: THE CHALLENGE OF EMAIL ARCHIVING

Andrea Goethals **Wendy Gogel**

Office for Information Systems
Harvard University Library
90 Mt. Auburn St.,
Cambridge MA 02138 USA

ABSTRACT

Because of the historical value of email in the late 20th and 21st centuries, Harvard University Libraries began planning for an email archiving project in early 2007. A working group comprised of University archivists, curators, records managers, librarians and technologists studied the problem and recommended the undertaking of a pilot email archiving project at the University Library. This two-year pilot would implement a system for ingest, processing, preservation, and eventual end user delivery of email, in anticipation of it becoming an ongoing central service at the University after the pilot. This paper describes some of the unexpected challenges encountered during the pilot project and how they were addressed by design decisions. Key challenges included the requirement to design the system so that it could handle other types of born digital content in the future, and the effect of archiving email with sensitive data to Harvard's preservation repository, the Digital Repository Service (DRS).

1. INTRODUCTION

1.1. Value of Email

Recognizing the potential long term value of email content to Harvard's research collections, the Harvard University Library charged a working group of University archivists, curators, records managers, librarians and technologists to describe the challenges of collecting, managing and archiving email at the University and to make recommendations for possible action. The group's March 2008 report highlighted email as an essential, yet missing part of our collections, and recommended that the University Library undertake a pilot project to build a system that would enable ingest, management, basic preservation, and also pave the way for access to email. The report emphasized the administrative, historical and legal value of email to the managers of manuscript repositories, archival programs and University records at Harvard. They also recommended that we identify critical policy and curatorial issues and address any legal or security concerns.

It is now widely recognized that email represents a slice of the late 20th and early 21st centuries (so far) that will be significant to historical research in the future. For our curators[1], collecting email represents a continuation of their traditional collecting in the categories of organizational records and personal papers. Since these records do not directly replace any single genre of analog content[2], their importance to future research only begins with their function as correspondence. We now appreciate email as a complex communications package that may contain unique primary source material; often serves as the document of record for business activities, decisions and outcomes; and is critical to the preservation of recent scholarly communications. The package includes the headers, message bodies, and attachments.

1.2. The Pilot Begins

For the pilot project, we were given funding for one developer for two years. We modeled our pilot project on the Libraries' successful first born-digital project, which resulted in the establishment of a central service at the University for archiving web resources - the Web Archiving Collection Service (WAX). WAX also started as a two-year pilot project, and entailed building a system to collect, process and archive resources to the University's preservation repository - the Digital Repository Service (DRS). Like WAX, the email archiving project would be managed and developed by the Library's Office for Information Systems (OIS), and would involve curators from throughout the University.

[1] In the rest of this paper the term curator is used to refer to any Harvard collection manager including archivists, librarians, museum and special collections curators and records managers.

[2] Note that the author of [7] and [8] changed positions where [7] describes email as an equivalent to correspondence and [8] notes that it has no parallel in the analog world. Our thinking during the pilot evolved along the same path.

Staff from three University repositories - the University Archives, Schlesinger Library, and Countway Library would partner with OIS to work closely on functional requirements and to supply email collections for testing. To address the legal and security considerations, we would consult the University's Office of the General Council and the University's Technology Security Officer.

Early on, the team that was charged with implementing the pilot recognized that the challenges we were confronting from transferring data of unknown formats, through identifying and securing sensitive data, to providing authority control to manage the variations on people's names, email addresses and institutional affiliations applied also to the broader, pressing need at the University to manage all born digital content. The curators assured us that all of these issues were not new to the field, but that they simply needed new tools and work flows to manage collections that are increasingly composed of a hybrid of analog and digital content. We pledged to use an architecture that would be flexible enough to expand to other oncoming born digital content. Although the focus of the pilot project is on content that has been selected for its long term value and therefore requires deposit to our preservation repository, we envision that in the future, the central infrastructure will also need to support the temporary storage of email and other born digital content as part of the University's records management schedule.

Notably our charge from the working group did not include delivery to end users as a requirement for the pilot. To support the research and teaching missions of the University, we will eventually need to provide a user interface for online delivery of email to end users. However, it was recognized from the beginning that access issues would be too complex to address in the pilot time frame and therefore would need to be addressed in the future, after the important first steps of collecting and preserving the email. However, the pilot will need to provide a mechanism for curators to provide mediated access to the email collections for researchers and for legal discovery. In anticipation of future end user delivery, we are defining the requirements for rights management that would enable automated access restrictions to a larger audience, and would continue to support curator-mediated access to the collections.

1.3. Nature of Email

Because of the pervasive use of email, at first glance the special challenges it poses for preservation are easily overlooked. In the course of conducting this pilot, four primary challenges due to the nature of email were identified: the diversity of mail client formats, the overly-flexible structure and composition of email messages, the tendency for email to contain sensitive information, and the volume of messages typically contained in individual email accounts. Secondary challenges included the tendency for email to have viruses or spam content, and the presence of duplicate attachments within email accounts and collections.

Although the format of exchanging email messages has been formally standardized through the RFC mechanism[3], the format for storing email messages has not been standardized. The storage format, including the directory structure, packaging format and location of attachments, is left up to the developers of email clients to decide. For this reason, mail clients vary in the way that they organize content, so the particular email client software has to be taken into account when preparing email for preservation.

There are also differences among email messages that aren't related to the originating mail client. Email messages can contain message bodies in text format, HTML, or both. Technically the message bodies can be in any format, but because mail clients need to display message bodies to receiving parties, in practice, message bodies have been limited to text and HTML formats. Messages can contain attachments in any format, and can contain in-line images within HTML message bodies. Some email messages do not have message bodies - as is the case when an individual sends an email that only has attachments. All of this variation has to be taken into account when processing, indexing, displaying and packaging email for preservation.

Individuals often use the same accounts for private and business correspondence. As Clifford Lynch, Executive Director of the Coalition for Networked Information put it, "email mixes the personal and professional in an intractable hodgepodge."[4] It can be difficult to impossible to separate, especially given the quantity of email most of us have. In addition, email is considered by most a private correspondence that will never be seen by anyone other than the original receiving parties. For example, Harvard curators have acquired email in which credit card numbers have been passed, and in which private health matters have been discussed. Email is the first content likely to contain sensitive information that will be ingested into Harvard's preservation repository. As the pilot progressed, we came to the realization that the sensitive nature of email would require us to rethink and redesign our repository infrastructure.

2. PRIOR WORK

Whenever OIS begins a new large project, we always review the larger landscape for prior and current initiatives that can inform our work. About 10 years ago

[3]See RFC-5322 Internet Message Format, and the related MIME Document Series (especially RFC-2045 MIME Part One: Format of Internet Message Bodies).

[4]CNI Conversations, March 10, 2010.

there was a burst of research and projects focused on email archiving and preservation. This work primarily came out of various city, state and national archives. One of the earliest of these projects, the DAVID project, was conducted by the Antwerp City Archives from 1999-2003. This project exposed many of the legal and privacy-related challenges of email archiving, and argued that email archiving solutions need to include clear policies and procedures as well as technical solutions. They chose XML as the long-term storage format for email and developed a simple XML schema for storing the message body and metadata about a single email [2].

Many other projects have also chosen XML for the normalization format for email [3][4][6]. The National Archives of Australia (NAA) created Xena, an open source format conversion tool that can convert email in three formats to an XML format. Some authors [7] conclude that text may also be a suitable long-term storage format for email. Other formats, such as HTML and PDF were considered by some but ruled out for various reasons, including the loss of significant characteristics of email or an incompatibility with search and index technologies.

Recently there were a couple of high-profile email archiving projects, also conducted by archives. The Collaborative Electronic Records Project (CERP)[5], conducted by the Rockefeller Archive Center and the Smithsonian Institution Archives, ran from 2005-2008. The Preservation of Electronic Mail Collaboration Initiative (EMCAP)[6] was conducted by North Carolina State Archives, Pennsylvania State Archives, and the Kentucky Department of Libraries and Archives. The CERP and EMCAP projects wrote guidance on transferring and formatting email, software for acquiring and processing email, and they collaborated on an XML schema designed to hold email for an account.

In addition to the DAVID, NAA and CERP/EMCAP schemas, there have been other efforts to develop XML schemas for email for general use [1] [5] [10]. In the early phases of the pilot we analyzed each of these schemas. We have preliminarily chosen to use the CERP/EMCAP schema, because we think it strikes the right balance between fully supporting the complexities of email headers and structure with a welcome lack of manipulation of the message bodies and attachments. Unlike most of the other schemas, it uses generic <Header> elements to store the names and values of the message headers. The advantage of this approach is that it can accommodate unanticipated headers, for example custom headers added by client systems, or those that will be added to future revisions of the email RFCs. It can support multiple message bodies per email, including HTML, and pointers to externally-stored attachments. They also have a separate schema for wrapping base64-encoded attachments; however we will likely decode attachments and store them in their original formats. While the CERP/EMCAP schema is designed to contain all the email messages for an account, we anticipate that it will work equally well at storing a single email message, which is how we intend to use it.

[5]See <http://siarchives.si.edu/cerp/index.htm>

[6]See <http://www.records.ncdcr.gov/emailpreservation/>

3. KEY REQUIREMENTS

To begin gathering functional requirements from our curatorial partners, we walked through several potential work flows with them. The scenarios covered the likely life cycle of email including the activities of email creators, data transfer to us, processing by the curators and then preservation in our Digital Repository Service (DRS). For the pilot project, we knew that we could not control or automate every step of the work flow and began working with the developer and other architects to refine the project scope.

A number of interesting challenges arose during this process. First, we were warned that a veritable tsunami of born-digital content was headed our way and that email would be only one of the great waves. Given the rate at which we all produce digital content, this was readily understood. Since the tsunami would include genres besides email, we are challenging ourselves to build a system that can grow and be generalized for other genres in the future. In light of the expected great wave of email, we recognized the likelihood that there would not be sufficient resources to process all of the collections at any depth. This led to the requirement to support mass transfer of content to the DRS with minimal manual processing, so that first and foremost, it would be safely and securely stored. It was determined, however, that the value of some collections would merit item-level processing of individual email messages and that this too would need to be supported. In keeping with traditional practices, curators would need to be able to return to collections that were only minimally processed and engage in more in-depth processing at a later time. This might occur because resources become available, to answer a research request, or because the value of the content has been newly assessed. This requirement - to enable processing the collection after it is transferred to the digital repository - is different than any other email archiving project we know of.

Second, we confronted new and very stringent requirements because of the potential for email to contain sensitive data as mentioned above. Although the laws vary from country to country and even within regions (or states in the U.S.), all email archiving projects need to confront security requirements to comply with laws at multiple levels of governance as well as local security policies and practices. At Harvard, this would influence the design of the new system as

well as have a profound impact on our existing infrastructure. Our email collections will likely include data that is defined by Harvard's enterprise security policy as High Risk Confidential Information (HRCI) and protected by Massachusetts State Law regarding personal information (201 CMR 17.00). Both are meant to safeguard personal information against unauthorized access or misuse and they generally cover a person's name in combination with identification numbers (such as U.S. Social Security or state driver's license numbers) or financial account information. In addition, some data will be protected by United States federal laws such as the Family Educational Rights and Privacy Act (FERPA), and the Health Insurance Portability and Accountability Act (HIPAA).

In consulting the University's information technology security experts, we discovered that email would need to be encrypted any time it was transported over a network or stored on portable media such as tape. Any applications accessing the content would need to be on the University's more secure private network, not the public network used in our existing infrastructure. Unfortunately, our DRS architecture did not comply with these security requirements. We needed to re-architect our repository to be able to accept, manage and preserve the email content. We also learned from the curators that, because of the sensitive nature of some of the content, only authorized people within the specific Harvard unit that stewards the collection would be able to view the contents.

A third challenge, reflecting the current collecting practices of our curators, is that email will represent both Harvard and non-Harvard content and may be closed or open-ended collections. The first email content contributed to the pilot project will be new content for existing analog collections. Email will be collected for noted figures in academia, science, politics and the arts (some of whom are faculty) and for institutions and organizations in areas where the University already collects "papers." This content defines the requirement to accept email from multiple mail servers, both internal and external to Harvard, from multiple types and versions of email clients, and to accept content from active as well as inactive accounts. In at least one case, the curator collected old email on a hard drive and has since negotiated with the creator to receive email on an ongoing basis.

4. DESIGN

Out of our review of prior work and the curatorial and security requirements we began to design an email archiving system that could integrate with our existing central infrastructure for authentication, authorization, persistent naming, discovery, preservation, and management. In past projects, we were accustomed to making small or no alterations to our existing infrastructure to accommodate new types of content. We envisioned adding a front end application named EASi (Email Archiving System Interface) to our infrastructure that would be used to prepare and push email into the DRS for preservation. Initially this seemed even simpler than WAX, which also acts as a specialized ingest system for the DRS, because WAX includes a complex crawler system.

After learning about the tsunami of born-digital content heading our way, we re-envisioned EASi as a front-end that could eventually accept whole hard drives of mixed content for processing and archiving to the DRS. Email would just be the first genre supported by EASi. It would now stand for the *Electronic* Archiving System Interface—not the *Email* Archiving System Interface. The EASi software developer is designing it so that it could be extended to other genres of content, and in a modular way so that the processing tools could be reused in the DRS management application, which would allow curators to continue to process the email, even after it is stored in the DRS for preservation.

Because currently there isn't a central server that we can pull email from, we are using a push model to get email into EAS. The overall data flow begins with curators transferring email to a central storage location at OIS via sftp, where an EAS process will pick it up and import it into the system. The curators will then be able to process the email using the EASi web-based interface, which will allow them to search, browse and read the email and attachments. They will be able to organize the email into collections, add rights and access restriction metadata, associate email addresses with people and organizations, delete email and/or attachments, and select content to send to the DRS. For this selected content, an EAS process will automatically prepare, package, transfer and load it into the DRS. After the content is in the DRS, curators will be able to continue to manage the email along with all their other DRS content (images, audio, etc.) using the web-based DRS management interface. As requested by the collection managers, this work flow is designed to support multiple levels of processing. Curators will have the option to do minimal processing up-front before pushing it into the DRS, knowing that they will be able to do further processing of the content later using the DRS management interface. Alternatively the system will also support more in-depth processing before pushing it into the DRS, if the content warrants the extra up-front effort.

When email is imported into EAS, the content is put through a number of automated processes. The first process converts the email to an RFC-282 Internet Message Format [9] using Emailchemy [11]. Table 1 lists the mail formats that EAS will support in the pilot phase because they are supported by Emailchemy. All the mail clients used by the pilot collections are supported by this list except for one obsolete DOS-based client called cc:Mail. We are still investigating whether

we can support this client using other tools. After format normalization the content is virus-checked, scanned for some forms of high-risk confidential information (initially just credit card and social security numbers), and scanned for spam. Finally the email content is parsed, metadata is extracted, and the metadata and content are indexed.

Email client software name and version	
AppleMail *	Outlook Express for Windows 4-6
AOCE *	Outlook Express for Mac (Database file) 5
AOL for Windows *	Outlook Express for Mac (Messages file) 5
Entourage *	Outlook for Windows
Eudora for Mac *	PowerTalk *
Eudora for Windows *	QuickMail Pro for Mac *
Mac OS X Mail 1-4	QuickMail Pro for Windows *
Mailman 2	Thunderbird
Outlook Express for Mac 4	Yahoo
Outlook Express for Unix 4	

Table 1. Formats that will be supported by EAS import

In response to the security requirements mentioned earlier, OIS system administrators came up with two options for redesigning the DRS architecture. Essentially the first option was to treat all DRS content as having sensitive data; the other option was to segregate the content coming through EASi from all the other DRS content.

4.1. Option 1: Integrated Content

In this option the entire DRS storage system would be moved to the more secure, more expensive private network. The advantage would be that we could continue to use the existing tape and disk copy infrastructure. On the negative side, we would not be able to use NFS to access the DRS files for the delivery and management applications anymore because it would be a security hole. It would need to be replaced with an ssh filesystem, which isn't known to scale at this time. In addition, the DRS management applications would need to be altered to use HTTPS connections and all DRS curators would need to access them using purchased Harvard VPN clients, even if they didn't use EASi. Because the tape backups of the EASi content would need to be encrypted we would have to purchase a separate tape library along with a SUN-encrypted backup service. Lastly, there was the concern that the front end delivery systems, which would remain on the public network, could be used to break into the back end secure system. Clearly this option had a lot of disadvantages.

4.2. Option 2: Segregated Content

In this option we would have two separate DRS storage systems – one for the content that entered through EASi, and the other for the rest of the DRS content. We would put the EASi applications and storage system on the more secure, more expensive private network, and leave the rest of the DRS on the existing network. One disadvantage is that we would need to replicate many of our DRS applications on the two different networks. We would need to replicate the DRS ingest applications that EASi uses to package and load the content into the DRS, and the DRS management application. The secured instance of the management application would need to be accessed using a Harvard VPN client. Although the content would be segregated, we could make it appear integrated from the curators' perspective because the DRS management application would be able to access both sets of content, allowing them to search and manage any of their DRS content together in the same interface. Although this option didn't allow us to leverage our existing architecture to the extent we would have liked, this seemed the better of the two options, so we proceeded to implement these changes.

5. FUTURE WORK

Although our current plan is to segregate the content coming through EASi from all the other DRS content, in the future we are optimistic that we will be able to reintegrate them. OIS system administrators are monitoring upcoming storage solutions that would allow us to have a more integrated solution when we do our next large storage migration, expected to take place in a few years.

The pilot project did not include within its scope delivery of the email content to end users—this will need to be undertaken as a separate post-pilot project. Prior to having a delivery service in place, curators and archivists will be able to access copies stored in the DRS for themselves through the DRS management application. This will allow them to provide mediated access to the email content for researchers, or if needed, for legal discovery. A delivery service for the email will entail more than the technical work of developing the delivery service application. It will also require expanded rights management metadata in the DRS, an overall strategy for collecting email at the University, and policies governing the range of activities from collection through delivery.

As we considered each of our key challenges and addressed them within the limitations of the pilot project, inevitably we thought about what we could do

given enough resources. In response to the born digital tsunami expected by the curators, we envision developing an environment where curators could appraise and process incoming content in a temporary holding area until a decision about its disposition can be made. When it is determined that the content will be accessioned for long term preservation storage and access, it would be transferred to the DRS and described in one of the public catalogs. We would make another repository available for transitory content that needs to be held for a limited amount of time according to a records management schedule or for specific legal reasons. We also envision building a centralized vocabulary registry that could be used by all of the metadata services in our infrastructure to help curators with authority control of terms including the various versions of people and institutional names and email addresses found in email.

6. REFERENCES

[1] Borden, J. *An RDF XML mapping of RFC 822/MIME*, The Open Healthcare Group, 2001. http://www.openhealth.org/xmtp/

[2] Boudrez, F., Van den Eynde, S. *Archiving e-mail*. DAVID project, Stadsarchief Stad Antwerpen, 2002.

[3] Dutch National Archives. *E-mail-XML Demonstrator: Technical description*, Testbed Digitale Bewaring, 2002.

[4] Green, M., Soy, S., Gunn, S., Galloway, P. "Coming to TERM: Designing the Texas Email Repository Model", *D-Lib Magazine*, 8(9), 2002.

[5] Klyne, G. *An XML format for mail and other messages*, 2003. http://www.ninebynine.org/IETF/Messaging/draft-klyne-message-xml-00.txt

[6] Moore, R., Baru, C., Rajasekar, A., Ludaescher, B., Marciano, R., Wan, M., Schroeder, W., Gupta, A. "Collection-Based Persistent Digital Archives - Part 2", *D-Lib Magazine*, *6*(4), 2000.

[7] Pennock, M. "Curating E-Mails: A life-cycle approach to the management and preservation of e-mail messages", *DCC Digital Curation Manual*, 2006.

[8] Pennock, M. *Managing and Preserving E-mails*. Digital Curation Centre, UKOLN, 2006.

[9] Resnick, P. (ed). *RFC-2822 Internet Message Format*, The Internet Society, 2001.

[10] Warden, P. *An XML format for mail and other messages*, 2003. http://www.ninebynine.org/IETF/Messaging/draft-klyne-message-xml-00.txt

[11] Weird Kid Software. *Emailchemy - Convert, Export, Import, Migrate, Manage and Archive all your Email*, 2010. http://www.weirdkid.com/products/emailchemy/

PHAIDRA - A REPOSITORY-PROJECT OF THE UNIVERSITY OF VIENNA

Markus Höckner **Paolo Budroni**

University of Vienna
Dr.-Karl-Lueger-Ring 1
1010 Wien

ABSTRACT

Phaidra (Permanent Hosting, Archiving and Indexing of Digital Resources and Assets) is used as a long-term preservation system through the assignment of persistent identifiers (permanent links). The project was launched in 2006 and is the successful result of cooperation between the Vienna University Computer Center, the Center for Teaching and Learning and the Vienna University Library. At present, Phaidra contains about 60,000 objects, all of which are provided with structured metadata.

1. INTRODUCTION

Like any other organization, the University of Vienna preserves millions of born-digital documents in several databases and file systems. Researchers are faced with two major problems: firstly, when they need to share their digital assets, they usually solve this problem by exposing their data on the internet (e.g. via a website). In this case, special knowledge is required to secure such precious assets in order to secure them from public downloading of these objects.

Secondly, researchers must provide a solution for long-term preservation [7]. Today, we produce millions of megabytes of data every day and we pay almost no attention to the fact that this data may not be accessible or reusable in the next few years.

Often the preferred solution is to build a repository. As a consequence, many repositories containing different technologies and metadata were built worldwide. But is there some method or solution available that would enable these to interoperate and exchange data?

In this context, the problem of long-term preservation is often underestimated. Only experts deal with this problem and they try to attract the attention of the producers of digital data regarding access problems which could occur in the future.

The University of Vienna has thus decided to address this challenge and started its own university-wide project in 2006.

2. NON TECHNICAL INNOVATIVE APPROACH AND SOLUTIONS

2.1. The long term perspective

Phaidra stands for a long-term archiving and digital asset management system and enables employees in teaching, research and administration to save, document and archive digital data and resources over a long period of time. Data can be systematically collected, equipped with multilingual metadata, assigned various rights and made accessible worldwide and around the clock. The continuous citability allows the exact location and retrieval of prepared digital objects. Phaidra can be actively used by all staff and students of the University of Vienna (via mailbox or u:net account). The objects can be viewed worldwide. Phaidra was developed at the Vienna University Computer Center in cooperation with the Vienna University Library and Archive services and the Center for Teaching and Learning. The project management is located in the University Library.

2.2. Project organisation and history

The project Phaidra bears three acting bodies: the Advisory Board, the Project Management and the so called Pilots.

The Advisory Board, an inter-university group of experts, performs strategic functions and supervises the achievement of strategic goals. The Project Management (two employees), based at the Library and at the Computer Center (three developers), is the operative entity of the project. The pilots for this were formed by clustering the needs of a few larger customer groups, like faculties, huge scientific projects as well as the University Library itself. The common challenge was to offer a vision of how to build a digital asset management system able to respond to the special demands of every department, institute and individual working at the university. The responses to this challenge were very

important for the acceptance of the project by the potential user groups involved.

The project itself was approved for three years by the authority of the University of Vienna. Only after one year of developing the first release (April 2008) has been deployed. Two years later, in February 2010, version 2 of Phaidra has been presented.

In April 2010 the development of the project has been extended for another three years. One of the main reasons was that there is a big demand of such a system at the University of Vienna. A lot of projects and organizations show interest in such a system.

2.3. Relevant factors

Phaidra was expected to host all digital assets stemming from the fields of

- research,
- teaching,
- technology enhanced learning and
- management,

and therefore it now offers general search functionalities covering the full range of stored assets as well as faculty and project-specific areas.

Concerning the metadata, a part of the chosen solution (which previously always was and still is a fountain of interesting and profound discussions) is the structured metadata, collected in an individual metadata schema. This metadata can be used for almost every form of digital content produced at the University of Vienna and all other institutions that are using the system and various other aspects of the long-term preservation of data.

Legal issues are also a very complex subject. The chosen solution addressing such was the involvement in the decision-making progress of a legal consultant specialised in internet law and intellectual property rights. The participation of the labour union of the University was a crucial factor in identifying the appropriate solution for the use of terms to be applied in the Phaidra project.

The integration of Phaidra into other services of the University of Vienna is a major challenge. Especially the connection between Phaidra and further legacy databases and systems as e.g. Fronter are an essential step in the success of the project. Other connections, for example to the projects EOD[1], WHAV[2] and E-Theses[3], are established and the progress of integrating objects into Phaidra has started.

But there are also efforts of the university to centralize the numerous different storage systems. Main reasons for this step are centralization, financing and maintenance. As a consequence the number of objects in Phaidra will dramatically increase in the next few months.

2.4. Strength of Phaidra

- Unique features
  The University of Vienna is currently developing an open access policy to motivate researchers to store the intellectual output of the research activities in Phaidra and grant free access to it.
- Access rights
  All persons who have a contract of employment with the University of Vienna as well as all of our students are allowed to upload assets into Phaidra. Guest accounts are included and currently in use. The world is able to view (read-mode) and/or download the assets except in the case where the assets owner restricts access to specific groups of people, individuals or even fully hides the asset. The latter means that only the owner is able to access the object and to modify the metadata. No one is allowed to delete any object.
- Terms of Use
  The Terms of Use stipulate the duties and the rights firstly of the service provider (which is Phaidra) and secondly of the systems users: usage of log files, users commitment to correct conduct (e.g. maintaining awareness of copyright issues), allowance to delete illegal objects, and security issues. Special terms provide regulations in case of the establishment of groups coordinated by a Super-User which holds the maximum amount of rights.
- Licensing
  A person who uploads an object must choose one of six plus one licenses, otherwise they are not able to finalize the upload process. There are six creative commons licenses, the GNU license and one general license available. Finally, users have the option of not choosing any license but keeping all rights reserved.
- Formats
  A document (best practice) informs users about formats that are recommended in order to achieve best permanent digital preservation (see section 3).
- Ease of Use
  Several tutorials and guidelines have been developed to support target groups to properly use the system.
- Training and dissemination work
  In addition, workshops are offered monthly to train people how to use Phaidra. Members of the Phaidra Team are also organizing regular meetings at the faculty level (Phaidra Days), in order to develop the dissemination effort, reach

[1] E-Books on Demand
[2] Western Himalaya Archive Vienna
[3] University of Vienna archive for electronic theses

broad acceptance within the university and eventually enhance users' willingness to share learning objects and other materials.

- First-Level and Second-Level Support
  Customer-oriented service is carried out by a Customer Manager. The Help Desk of the University has been trained for the first-level support. Second-level support is provided through the technical development team. The service website[1] additionally offers extensive information about the system and respective services.
- Updates
  Updates are made once a month.
- Classification of Digital Objects
  Several subject-specific thesauri have already been implemented in order to support indexing.

[2]

3. THE SYSTEM PHAIDRA

As mentioned previously, Phaidra is a digital asset management system with long term preservation aspects[2]. A technical overview will be given in section 4. At this point, we will take a closer look at the system and its limitations.

Every repository stores a huge amount of different data. This data can contain pictures, audios and so on. But there is also the need to store more than one content into one object or sometimes no content at all. So Phaidra differs between three groups of objects:

1. Single-File-Object (one content datastream)
2. Container (multiple content datastreams)
3. Collection (no content - only members of the collection)

With the help of these three-object groups almost every content can be ingested into Phaidra. But these categories are not as precise as desired.

As a consequence, Phaidra differentiates between different object types. These types are Picture, Audio, Video, Document, Resource, Container, Collection, Book, Page, E-Paper and Asset. These eleven different object types are the different content models of Fedora. How these object types are used in Fedora will be described in section 4.1.

To be able to archive and present webpages, Physlets and so on Containers are used. Because of the fact that these type of content has numerous files the Container has been created. Using for these special type a Collection would mean to create numerous objects in the repository that make no sense.

The types "Book" and "Page" are special types in the repository which make it possible to produce online books. These books [3] can be viewed via browser (Phaidra Book Viewer) and if OCR data is available there is also the option of setting up a search. But these object types do not accept every file format due to long-term preservation constraints.

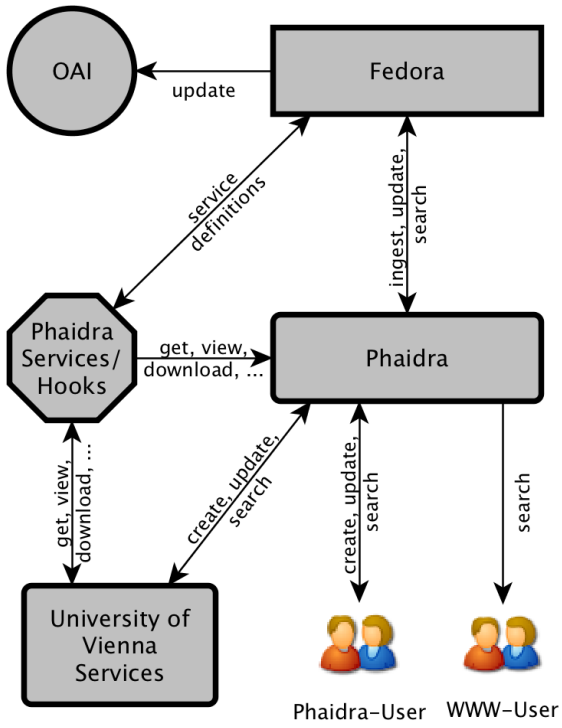


Figure 1. PHAIDRA

Phaidra is designed to migrate the content if it is necessary because in future some types of file formats will be detached by other formats. For this reason, not every content is designed for long-term preservation because of proprietary or inconsistent file formats. Because of this fact, Phaidra only allows a small subset of file formats for the different object types. For example, in the case of the type "Picture", the recommended format is TIFF; also allowed are JPEG and JPEG2000; all other formats are not allowed. Thus, if a user wants to upload a GIF image they must choose the Asset type because this special type allows every content type (at the web frontend, this type is called "Unknown"). So the option exists to upload every content into Phaidra but if the content is declared as an Asset, Phaidra will not take responsibility that it can be accessed in the future.

4. TECHNICAL ASPECTS

PHAIDRA has been implemented expecting 80.000 concurrent users. It can be divided into two parts.

[1] http://phaidraservice.univie.ac.at/
[2] http://phaidra.univie.ac.at/o:52318
[3] https://phaidra.univie.ac.at/o:19958 - Plinius Historia Naturalis - The oldest book of the University of Vienna

Firstly, there is the web frontend that allows members of the University of Vienna to create, update and search objects. Secondly, there is the well-known repository Fedora that is used for storing the objects with their metadata.

In the following sections, the connections between (see Figure 1) these two will be explained.

4.1. Fedora

This well-known open source repository is very often used in projects just like Phaidra [5] because it is very reliable and can be easily adapted to special demands [1]. It is implemented in Java and supports features like storing all types of multimedia and their metadata, an API for accessing the repository itself (SOAP and REST), provides RDF search and so on.

Phaidra employs a modified version of Fedora. The modifications to this version took place in different steps. One step was the need of hooks that should check if the submitted metadata is valid. Also a modified search in the repository has been implemented. But all of this was no problem because of the advantages of this repository mentioned before.

Because of the fact that Phaidra uses a modified Fedora the communication with Fedora-Commons/Users is very close. On the one hand because of some kind of bugs and on the other hand because of modifications that may be interesting for the community.

If you integrate an object into the repository, every object will receive a unique identifier. This identifier will never change and so Phaidra uses this identifier as a permanent identifier. For this reason, every object in the repository will be accessible under the same object ID as long as the repository exists.

Fedora uses a Content Model Architecture to differ between the different content that was integrated into the repository and present it to the requester. If you only want to retrieve the content as it is stored, there is no need of using the CMA because Fedora is using its CMA behind the scenes. This architecture is very important for repositories because it is the form of communication of Fedora with other systems.

So Fedora is able to differ between the different content that is stored in the repository. But not every item of content can be returned to the requester as it is saved, because the requester may not interpret the mime type of the content. So the content has to be converted.

Fedora itself is not able to convert content, but with the help of the CMA and the possibility of defining services, this problem has been solved.

These service definitions (see Figure 2) represent the ways Fedora communicates with other services, for example, converting the content from format X to Y. Every content model in Fedora has service definitions and tasks to do different jobs that can be defined by the administrator of the repository.

A service definition defines services and different operations. To carry out the requested job, you must also define service tasks because the service definition will not know how to interpret the service outside of Fedora. The defined service deployments can be linked to the different service definitions and so Fedora knows what to do if a certain operation is requested.

There are different ways of searching in Phaidra. First, you can search in the index with the help of SPARQL. Before Fedora 3.3, this method was sometimes very inefficient because it was very slow. Now, searching in the index is about ten times faster. For example, the OAI-PMH provider can be used to search for new created or updated objects.

In addition, a full-text search is also available. It is called GSearch and with the help of this, you are able to search in defined fields of the index. To perform the search, Lucene is used because it is quite fast and common. So if you configure Fedora to extract the fulltext of the PDFs during integration, you will then be able to search through all documents and receive a result of the matching hits.

Actually about 60.000 objects are in the repository available. These objects are saved on the main SAN of the University of Vienna and use about 2 Terabyte of space. Most of the objects are Picture (about 20.000) and Page (about 30.000) objects.

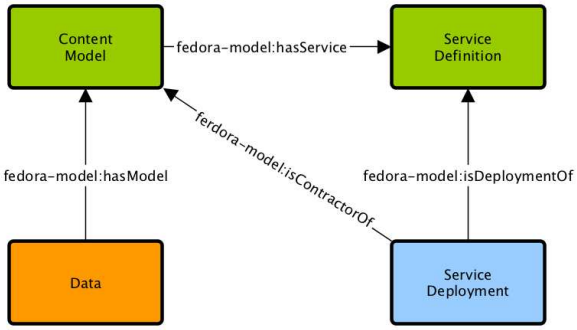


Figure 2. Fedora CMA Relationships

4.2. Phaidra

4.2.1. Phaidra Core

There are several interfaces for Fedora available just like Muradora and so on. But the University of Vienna decided to create an own interface. Main points for this decision were

- adapting the interface to the internal structure of the university,
- adapting the interface to existing ones,
- designing an "own" interface.

Because of the fact that most of the web applications at the University of Vienna are written in Perl the programming language of Phaidra Core is also Perl. So

until 2013 the Core will be developed in Perl. But the Phaidra Team traces the development of the other interfaces with great interest.

To assure scalability, extensibility, reusability, flexibility and reliability the web application framework Catalyst [6] is used.

Catalyst is a widespread framework for Perl and has a very strong community. Because of the fact that a lot of web applications work with Catalyst, there are a lot of plugins available for it. For this reason, the developers do not have to implement the web application from the scratch. They can then concentrate on other issues that may arise.

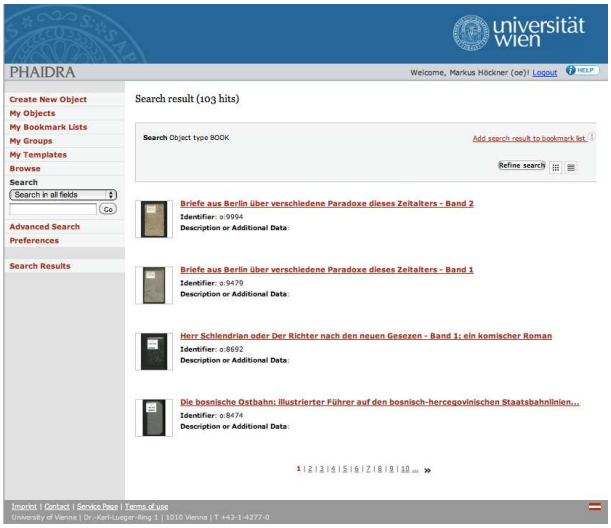


Figure 3. Phaidra Webfrontend

For integrating objects, updating the metadata or searching in the repository SOAP is used. Since Fedora 3, there is also the possibility of using REST, but when the project started this option was not available. So Phaidra must be connectable to a specific API-A or API-M method of Fedora and put the data into Fedora or retrieve it.

For character encoding, Phaidra uses UTF-8. Because Phaidra is able to handle almost every language and their lettering in the metadata and webfrontend. So there is no need for a coding or decoding location that would otherwise cost great performance.

The webfrontend (see Figure 3) is fully localized in German and English and soon there will be an Italian version available. When the web frontend was developed, the usability of the application was very important. So the application was designed as simple as possible and clearly structured. To fulfill these demands, new web technologies like AJAX are used.

But there is also the need for batch or automatic uploads. To carry these out, a special Phaidra API was developed by the Phaidra team. If a user wants to upload objects into Phaidra they do not have to use web frontend. The creator of the object is able to use this API to create and update objects in Phaidra. Also, a search method is available. And now this API is available for Perl and Java.

4.2.2. Phaidra Services/Hooks

The Services and Hooks of Phaidra are very important parts of the web application because they have to perform much work.

To be able to convert pictures, to view them in the browser, to download a PDF document or to play a video, other services are required as Fedora is not able to do this work. So not only the web frontend and the ingest of objects are main developing points, also the services are important. They are responsible for presenting the content in the appropriate way.

For example, a picture is uploaded into Phaidra and you would like to view it in your browser. There might be the problem that the uploaded content is a TIFF image. So without a browser plugin you are not able to view it. To prevent that every user is able to view this picture in the browser it has to be converted in a format that every browser can interpret, for example, JPEG.

So the first step is that Fedora recognizes that you need a JPEG image. For this, the CMA of Fedora is used. Fedora recognizes the object as a picture due to the CMA. With the help of the service definition, the system connects the object to the appropriate Phaidra Service. The system calls this service to convert the picture into JPEG format.

In Fedora, almost every kind of data can be stored. So there is no problem with saving metadata in an XML and adding this XML as a datastream to the object. The problem is that if you have a metadata schema, you must also define a structure for the XML as well as special vocabulary. To ensure that only valid XML datastreams are added to objects in Fedora, Phaidra has so-called Hooks. These Hooks are responsible for checking if the metadata is valid and the object has all of the required datastreams and service definitions. For this reason, these hooks are possibly the most relevant parts of Phaidra because they guarantee reliability and security.

5. UNIVERSITY OF VIENNA METADATA

Metadata is structured data about other data and fulfills a variety of tasks [3].

- identify objects worldwide;
- describing objects (e.g., author, creator, title, description);
- support of information retrieval and identification;
- describing the historical audit trail of an object and its provenience;
- grouping objects into collections;
- rights information, licenses and access permissions;
- technical information about the content;

- easier interchange of data between autonomous repositories;
- versioning of an object;

Because of this metadata, the objects are somehow self-documenting and prepared for long-term preservation. But most of the metadata must be created by humans. As a consequence, metadata costs a lot of money and time to preserve every object in the best way.

The metadata schema of the University of Vienna is a modified LOM schema. LOM is a standard by the IEEE [4] that is well known for describing and documenting learning objects in a repository. The standard describes the basic structure of the metadata, datatypes, list values and vocabularies. The need of such a standard is significant because of interoperability and long-term preservation.

The first version of Phaidra contained the LOM schema with some specific adaptions just like extending the vocabularies. After a certain time, faculties of the University of Vienna got in contact with the Phaidra team and the metadata working group. The reason for said contact was that they also had data to store but they needed specific metadata. As a consequence, the LOM schema had to be extended again. So the process of analyzing and adapting started again and still continues.

In the last two years, two major adaptations were made on the metadata schema of the University of Vienna. First, there was the need to store primary data. Especially the Institute History of Arts at the university asked for this new section. Thus, cooperation between Phaidra and this institute was initiated.

Phaidra also stores digital books and so book-specific metadata (e.g., publisher, publishing date) had to be included into the schema. In cooperation with the Library of the University of Vienna, these new metadata tags have been added and included.

To be able to offer Dublin Core metadata, the University of Vienna metadata schema has eight mandatory fields. So the user of Phaidra does not have to add Dublin Core metadata manually because it is extracted automatically from the metadata schema of the University of Vienna. The Dublin Core will be saved into the object as a datastream so that it is easily accessible.

6. FORECAST

The project development of Phaidra has been extended by the University of Vienna for the next three years because of the need of such a repository and its applications. The main step in the project will be to set up connections to other systems at the university and to migrate data into the repository.

Besides the improvement of the existing streaming services there is also a need for collaborative functions for the SuperUser and for a new ImageViewer. This new application will allow users to view images greater than one hundred megabytes via the WWW. The basic elements for this new application exist (Phaidra Imagemanipulator) but there is still much work to do.

Phaidra also participates at several Europe-wide projects. The two best known are TEMPUS and OPENAIRE.

TEMPUS (Trans-European Mobility Scheme for University Studies) is a European Union programme. The aim of the project is the modernization of higher education in countries surrounding the EU. Through this project the University of Vienna stays in touch with several Western Balkan universities.

OPENAIRE (Open Access Infrastructure for Research in Europe) deals with the problem of open access at a European level. The project currently has 38 partners from 27 European countries.

One of the oldest universities of the world, the University of Padua, has contacted the University of Vienna due to the need of a repository. They from University of Padua took a closer look at Phaidra and decided to join the project. So Phaidra has been installed there at the beginning of May 2010.

Also, national universities and institutions have contacted the Phaidra group. Until now, three instances of Phaidra are planned, one for the University of Applied Arts Vienna, one for the University of Music and Performing Arts Graz and another for the Austrian Science Board. Phaidra is also in contact with two more Austrian universities.

Since the end of 2009, Phaidra is also a part of the project Europeana. The aim of this project is to make digital content efficient and quickly accessible at a European level. The first objects from Phaidra have been successfully transferred to Europeana. To enable this, the OAI-PMH provider of Fedora is used.

So the development of Phaidra is not finished yet. With the help of other resources, the project will grow in the next few years. Also, sharing of know-how with other universities and projects will help in the development Phaidra.

7. REFERENCES

[1] Bainbridge D. AND Witten I.H. ”A fedora librarian interface”, *Proceedings of the 8th ACM/IEEE-CS joint conference on Digital libraries* , Pittsburgh, U.S.A, 2008.

[2] Budroni P. ”Manifest zur Bildung einer Matrix”, *Mitteilungen der Vereinigung Österreichischer Bibliothekarinnen und Bibliothekare 63 (2010*) Nr. 1/2 , Bregenz, Austria, 2010

[3] Gladney, H.M. Preserving Digital Information. Springer, Berlin-Heidelberg, 2007.

[4] Institute of Electrical and Electronics Engineers, Inc. Draft Standard for Learning Object Metadata. New York, 2002.

[5] Kumar A., Saigal R., Chavez R. AND Schwertner N. ”Architecting an extensible digital repository”, *Proceedings of the 4th ACM/IEEE-CS joint conference on Digital libraries*, Tuscon, U.S.A, 2004.

[6] Rockway, J. Catalyst - Accelerating Perl Web Application Development. Packt Publishing, Birmingham - Mumbai, 2007.

[7] Waugh A., Wilkinson R., Hills B. AND Dell’oro J. ”Preserving digital information forever”, *Proceedings of the fifth ACM conference on Digital libraries* , San Antonio, U.S.A, 2000.

Session 4a: Trusted Repositories

BECOMING A CERTIFIED TRUSTWORTHY DIGITAL REPOSITORY: THE PORTICO EXPERIENCE

Amy Kirchhoff **Eileen Fenton** **Stephanie Orphan** **Sheila Morrissey**

Portico
100 Campus Drive, Suite 100
Princeton, NJ 08540

ABSTRACT

The scholarly community's dependence on electronic resources is rapidly increasing and those electronic resources are increasingly preserved in digital repositories or other preservation services. Whether locally hosted at libraries, collaboratively hosted between institutions, or externally hosted by a third party, one method for these digital repositories to take to assure themselves and their communities of their soundness is to be audited and certified by impartial organizations. Such independent organizations with staff experienced in executing audits and certifications can represent the interests of the academic community. Such staff will have the time and skills required to perform a thorough review of the methodologies and policies of each digital repository.

Over the course of 2009, the Center for Research Libraries (CRL) audited Portico, a third party preservation service. At the conclusion of the audit, CRL certified Portico as a trustworthy digital repository. The audit was a lengthy, productive experience for Portico. We share the experience here both to impart the depth of the audit and to inform other organizations of what steps might be involved should they choose to be audited and certified.

1. INTRODUCTION

Over the course of 2009, the Center for Research Libraries (CRL) audited the Portico preservation service. The audit formally concluded in January 2010, when CRL certified Portico as a trustworthy digital repository. Portico is the first preservation service so certified by CRL.

"The Center for Research Libraries (CRL) conducted a preservation audit of Portico (www.portico.org) between April and October 2009 and, based on that audit, has certified Portico as a trustworthy digital repository. CRL found that Portico's services and operations basically conform to the requirements for a trusted digital repository. The CRL Certification Advisory Panel has concluded that the practices and services described in Portico's public communications and published documentation are generally sound and appropriate to both the content being archived and the needs of the CRL community. Moreover the CRL Certification Advisory Panel expects that in the future, Portico will continue to be able to deliver content that is understandable and usable by its designated user community." [1]

Portico (www.portico.org) is a not-for-profit digital preservation service providing a permanent archive of electronic journals, books, and other scholarly content. Portico is a service of ITHAKA, a not-for-profit organization dedicated to helping the academic community use digital technologies to preserve the scholarly record and to advance research and teaching in sustainable ways. In May 2010, there were nearly 15 million articles and 2,000 e-books preserved in the Portico archive with over 10,000 journals, 30,000 books, and 10 collections of digitized historical content committed to the archive. We anticipate that an additional 1.5 to 2 million articles, tens of thousands of e-books, and several d-collections (digitized historical collections, such as historical newspapers) will be preserved in the archive every year.

CRL (www.crl.edu) is an international consortium of university, college, and independent research libraries.

The CRL audit of Portico extended through ten months from April 2009 to January 2010. It was the first preservation audit Portico has undergone. Ultimately, this audit was a collaborative and productive learning experience.

As an element of certification, CRL assigned Portico levels of certification in three categories: organizational infrastructure; digital object management; and technologies, technical infrastructure, and security. Portico's score for each category is given below in Table 1. (The numeric rating is based on a scale of 1 through 5, with 5 being the highest level, and 1 being the minimum certifiable level.)

Category	Portico Score
Organizational Infrastructure	3
Digital Object Management	4
Technologies, Technical Infrastructure, Security	4

Table 1. Portico Certification Scores

In addition to the formal scoring, Portico and CRL agreed that over time Portico would address some of the concerns CRL highlighted in the written audit report1 and in informal discussions with Portico (for example, improving the Portico roles and responsibilities documentation).

Portico benefitted from the audit in practical and tangible ways. Our preparation for the audit, which included collecting and updating documentation, made it easy to provide this documentation to other parties subsequent to the audit. The most significant benefit is the assurance regarding the viability, the integrity, and the effectiveness of our preservation approach that only such a comprehensive, objective, third-party review can provide.

2. REASONS TO BE AUDITED

An audit is "an evaluation of a person, organization, system, process, enterprise, project or product" [5] and certification is "the confirmation of certain characteristics of an object, person, or organization … this confirmation is often … provided by some form of external review, education, or assessment." [6] The CRL preservation audit and subsequent certification of Portico as a trustworthy digital repository was just that, an external review and evaluation of Portico.

The yearly statistics produced by the Association of Research Libraries (ARL) show that every year the scholarly community becomes more dependent on electronic content (see Figure 1). Indeed, by 2008 the ARL institutions were spending over 50% of their library materials expenditures on electronic resources – resources that by their very electronic nature are not preserved on the shelves of the library itself.

Portico preserves an ever growing portion of these digital resources the scholarly community relies upon, and as such we felt it was imperative that we undergo a formal third party assessment and certification process to assure ourselves, the ITHAKA board, and, most importantly, the scholarly community, that our preservation methodology, processes, and archive will secure the long-term preservation of the content in our care.

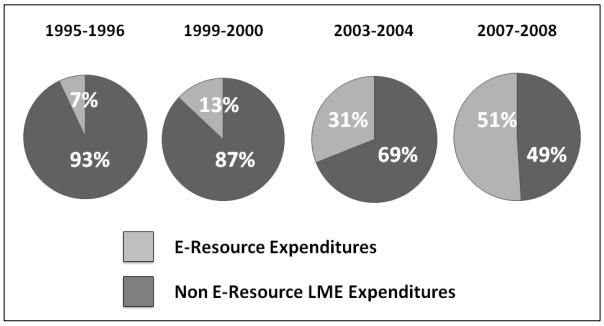


Figure 1. ARL E-Resources Expenditures2

3. AUDIT METHODOLOGY

CRL based its audit process on the Trustworthy Repositories Audit & Certification: Criteria and Checklist (TRAC), as well as other inputs of interest to the CRL community. These inputs included *"metrics developed by CRL on the basis of its analyses of digital repositories. CRL conducted its audit with reference to generally accepted best practices in the management of digital systems; the interests of its community of research libraries; and the practices and needs of scholarly researchers in the humanities, sciences and social sciences in the United States and Canada. The purpose of the audit was to obtain reasonable assurance that Portico provides, and is likely to continue to provide, services adequate to those needs without material flaws or defects and as described in Portico's public disclosures."* [1]

TRAC is a standard that was developed by experts within the digital preservation community. Its goal is to identify the criteria that define a trustworthy digital repository. It is important to be aware that TRAC and other digital repository audit methodologies, such as the Digital Repository Audit Method Based on Risk Assessment (DRAMBORA), are designed to evaluate a repository against its own claims, not against a single standard set of measurements. *"At its most basic level an audit should assess whether a repository can meet its stated commitments–is it doing what it says it is doing?–and the criteria have to be seen within the contexts of the special archiving tasks of the repository."* [2] With such a focus on the context of the specific repository being evaluated against TRAC, two repositories with very different levels of documentation, and indeed with very different kinds of preservation goals, service level models, and guarantees, could both be certified, if the level of documentation at each repository supports that repository's individual purpose and public statements.

The CRL audit team consisted of two full-time CRL staff members and one CRL technical consultant.

[1]http://www.crl.edu/sites/default/files/attachments/pages/CRL%20Report%20on%20Portico%20Audit%202010.pdf

[2]This chart is created from the publicly available figures in the ARL annual statistics at http://www.arl.org/stats/annualsurveys/arlstats/index.shtml

Guidance and advice on areas of concentration for all CRL digital repository preservation audits is provided to the CRL audit team by the CRL certification advisory panel, which represents the CRL membership and *"its community of research libraries and the practices and needs of scholarly researchers in the humanities, sciences and social sciences in the United States and Canada."* [1] The CRL certification advisory panel includes leaders in collection development, preservation, and information technology.

At Portico we made several important decisions early on in the audit process: 1) we agreed it was important to ensure that the CRL audit team understood our preservation philosophy, policies, and workflow, and 2) we would establish a primary contact for CRL throughout the process. The Portico archive service product manager, Amy Kirchhoff, coordinated the internal process and communicated externally with the CRL audit team, while many staff members of Portico and ITHAKA were involved in the audit process. In particular, the Portico senior research developer, Sheila Morrissey, and publisher content coordinator, Stephanie Orphan, were heavily involved in audit preparations. CRL and Portico collaborated on the development of the timeline and logistics for the audit process. Over the course of several conversations, we worked together to identify what documents would be required.

Portico gathered documentation and expertise from all parts of the organization and provided CRL with five subject based portfolios of documentation. To aid this portfolio creation, we developed an internal document cross-referencing nearly all of Portico's documentation to the TRAC criteria. Shortly after receiving the documentation from Portico, the CRL audit team visited the Portico New Jersey office to witness and audit the steps Portico takes in its preservation process. Following the site visit, there was an ongoing dialogue between Portico and the CRL team as we worked to address their questions about our preservation process, policies and documentation. While the audit itself was quite rigorous, it was a productive and collaborative process.

3.1. Documentation

As with virtually all kinds of audits, the CRL digital repository assessment requires the repository to provide evidence to demonstrate how it meets the audit criteria. This evidence-based methodology is intrinsic to TRAC, *"in particular, appropriate documentation of all steps permits auditors to evaluate the digital long-term repository as a whole"* [2] and DRAMBORA, *"a range of evidence expectations are described within the audit tool, reflecting a belief that organizations must be able to demonstrate their ability to effectively manage their risks."* [3]

In support of this evidence-based methodology, we spent several months identifying documentation we had already written and cross-referencing it to TRAC. Before the site visit, Portico provided the CRL audit team with 1,225 pages of documentation organized into five portfolios:

- **Organization**: including items such as organizational charts, meeting notes, financial statements, documentation of surveys, and sample email conversations with participants
- **Policy**: including all Portico preservation policies
- **System Architecture and Content Model**: including several introductory presentations, and content model & information architecture documentation
- **Operations and Systems Development & Maintenance**: including content manifests, illustrative documents from Portico trigger events and instances of post-cancellation access, sampling of minutes from the weekly technology & operations meetings, documentation for major systems changes, Portico disaster recovery plan, documentation of the results of retrievals from backup, support contracts with external vendors, receipts for payment of cloud storage service fees, and documentation about fixity verification processes, including recovery in case of errors found on disk
- **Archive Interfaces**: including user and business requirements for the audit and access interfaces to the Portico preserved content and documentation about planned enhancements to the auditor interface

For these portfolios, Portico staff collected previously written documentation and reproduced that documentation in image form. In order to provide context to each document, Portico wrote introductions to precede most documents. We completed significant writing for the audit in the area of policies–many of Portico's policies were encoded in training classes and operational procedures (which were also provided to the CRL audit team). Preparing for the audit created an opportunity for us to consolidate our understood "policies" into formal policy documents.

After receiving the portfolios of Portico documentation, visiting Portico on-site, receiving sample articles exported from the Portico archive, and reviewing all of the information gathered throughout the audit process, the CRL audit team requested additional documentation from Portico, including:

- Samples of the "Portico Modification to Original Submission Information Packages or Portico Archival Units Form"–a document Portico uses for tracking purposes when it is necessary to modify content outside of the standard ingest workflow, for example if prior to ingest Portico will be replacing corrupted content with corrected content as provided by the publisher.

- Sample format action plans (format action plans are documents that describe how an organization will address the preservation needs of specific file formats) and turn over documents (which specify the format action plans for publisher-specific XML and SGML formats and publisher-specific packaging schemes).
- Lists of formats and file types accepted into the archive and any formats and file types not accepted. Portico accepts all file formats into the Archive and provided the CRL team with a list of all formats in the archive (files in the Portico archive are assigned a preservation level determined by the tools available to support the file format and the commitments made to the specific content (e.g. well-formed PDF files associated with e-journal articles are fully preserved, whereas ill-formed PDF files or executable applications are byte preserved)—as file format tool sets improve over time, the preservation levels assigned to specific files will be adjusted.)
- Brochures designed for library and publisher outreach, provided as PDFs.
- Example license agreements as exported from the archive.
- Relevant technical certifications earned by Portico.
- Documentation of any hardware and software changes. This information is encoded in the event records in the archival information packages preserved in the archive.
- Budgets and expense/revenue statements for 2005-2009.
- Sample communication to publishers regarding status of their content. Twice a year, Portico provides publisher participants with a report that includes general information about Portico status and specific information about that publisher's content in the archive.
- A sample publisher agreement annex in spreadsheet form. This document lists what content is committed to the Portico archive.
- An explanation of the process used to produce library-specific holdings comparison reports– these reports compare the holdings of the Portico archive to those of a specific library or portion of a library's collection.

3.2. Beyond Documentation

In addition to producing the documentation portfolios and providing additional documents on request, Portico engaged with CRL through numerous phone and e-mail conversations. While Portico and CRL had a number of conversations about audit logistics, the majority of the conversations were initiated by the CRL audit team as questions arose during their review of Portico-provided documentation and sample articles. Many of these questions required responses rich with information and we appreciated the opportunity to clarify Portico policies and practices.

We received general technical questions from the CRL audit team, including questions about the Portico information architecture, replication policies, and bit corruption tolerance. (Portico has a zero tolerance policy, which is not documented separately, but is reflected in the fixity verification documentation.) The CRL audit team reviewed the sample articles in depth, compared them to the content model documentation we provided, and developed a variety of article-specific questions about identifiers and other required (or not) descriptive metadata, article presentation for delivery, and content transformation.

The CRL team was quite interested in exploring and testing retrievals from the Portico archive. In order to address this concern, we explained that we frequently export content from the archive including regular exports to the archive replicas, the delivery site, and (at the time of the audit and in accordance with existing publisher agreements) to the Library of Congress. In addition, we perform a number of one-off exports to our participating publishers.

The CRL audit team was also particularly interested in the Portico holdings and ways for the community to gain detailed information about the specific contents of the Portico archive. We discussed tools such as: the audit web interface through which librarians and publishers may review archived content, the Portico holdings comparison tool that compares a library's holdings to the Portico archive, and the detailed Portico holdings lists.

CRL also had questions about the business and technical logistics of providing post-cancellation access, a service that Portico provides to participating publishers on an opt-in basis. The CRL team also inquired whether Portico receives DTDs and schemas from publishers and whether they are placed in the archive and we confirmed that these materials are received and preserved in the archive.

3.3. Audit Timeline

Portico was involved in audit preparation and the actual audit for approximately 16 months (from the fall of 2008 through January 2010), although the audit itself extended over 10 months.

Winter 2008-2009–During the early winter of 2008, Portico and CRL held initial discussions about the proposed timeline for the audit.

Spring 2009–We began the process of identifying and collating existing documentation from a variety of departments, including finance, human resources, legal, information technology, content management, user

support, delivery, publisher relations, outreach, and operations. This documentation was distributed across many systems, including Talisma (a contact management system), SVN (a version control system), JIRA (a bug, issue, and project tracking system), the Portico intranet, a Wiki, shared drives, web servers, the public website, local drives, and email accounts. Portico also began work on a TRAC self-report documenting to what degree we met the 84 criterion in TRAC and describing the documents available to support our assessment. (This TRAC self-report is available on the Portico website in the Archive Certification area.) [3]

CRL announced the launch of the audit of Portico in March 2009 and in April, Portico submitted the TRAC self-report to CRL.

Summer 2009–Portico developed a policy document template and policy approval framework and began to document existing policies using the new template. We continued to create the five portfolios of documentation. In May, Portico and CRL agreed to the logistics of the audit and we learned who at CRL would be on the audit team and how the team would interact with the CRL certification advisory panel. We provided the CRL audit team with access to the Portico auditor website and received the schedule of documentation from CRL. In July, Portico and CRL finalized the agreement guiding the audit process. Portico also provided references for third parties that received data exports from Portico. Portico submitted the first portfolio of documentation, the organizational portfolio, to CRL on August 4th.. The four additional portfolios were submitted on August 13th.

On August 19th, the CRL audit team visited the Portico office in Princeton, New Jersey. As the CRL team was particularly interested in observing staff perform their normal activities, we arranged for the team to "follow" the content as it moved from one Portico unit to another. We started the day by attending the daily meeting between the technology and operations groups. Next we took the CRL team to talk with the publisher content coordinator who kicks off the business and analysis processes that begin after a publisher has signed a preservation license agreement. Next the CRL team spoke with the staff that develops publisher-specific tools that transform content to archival formats. The CRL team then spoke with members of the Portico systems team and attended a release coordination meeting. To end the visit, the CRL team met the Portico ingest team, where they witnessed the process of ingesting content into the archive and resolving problems with the content during the transformation process.

After the site visit, Portico staff wrote software to export the 200 sample articles requested by CRL from the archive and build a navigable set of HTML pages that would allow the CRL team to review the entirety of the archival information package for each article, including the archival metadata file (the current Portico auditor interface does not provide access to the archival metadata file or the publisher's original SGML or XML files). Portico made this set of pages available to the CRL via an FTP site, where we also made available the Portico tool registry, file format registry, business data objects (a database that maps Portico publishers to their titles, used for collection management purposes), and a set of 20 submission information packages (a submission information package is content as provided to Portico by the publishers before any archival processing).

Fall 2009–Portico and CRL interacted extensively and Portico provided additional documentation as requested (including job descriptions and additional financial information). In October 2009, we coordinated a conference call between the CRL audit team and the Library of Congress to allow the CRL team to learn from the Library of Congress about their experiences developing an export process with Portico and managing the receipt of content exported from the Portico archive.

Winter 2009-2010–Portico received the draft report from the CRL audit team and offered comments. In January 2010, CRL released the final audit findings, initially sharing the results with CRL members and then to the broader community.

Spring 2010–Portico and CRL continue to have conversations about areas of particular interest to CRL or in response to questions raised by the CRL membership.

3.4. Audit Costs

Over the course of the 16 months during which Portico was engaged in the audit process, many staff participated, including staff from library outreach, publisher outreach, legal, finance, user services, operations, and development. The Portico Archive Service Product Manager invested the most time, approximately four months of work. Combined, other staff contributed another four months of work. This staff cost was funded out of Portico's operating budget. Ongoing communications with CRL and the regular updates will also be funded out of Portico's operating budget. Portico will integrate addressing the concerns raised during the course of the audit into day-to-day operations. We believe the regular updates that must occur every two years will require significantly less staff time than the initial process.

3.5. Ongoing Audit Activities

The CRL report on Portico audit findings outlines concerns the CRL had on 12 of the 84 TRAC criteria [1]. The CRL team provided Portico with additional comments by email and phone. The concerns range from documentation discrepancies (e.g., discrepancies found between Portico job descriptions and Portico policy

[3] http://www.portico.org/digital-preservation/wp-content/uploads/2009/10/CRL-Audit-Portico.FINAL_.pdf

documentation) to very specific requests for a software and hardware patch register to more general concerns about usability. Portico is developing a road map that will allow us to address these issues over time. We remain in contact with CRL on areas of mutual interest (for example, how to share holdings information).

As appropriate, Portico has already addressed some issues identified in the CRL report.

The CRL report identified concerns with the opaqueness of the Portico holdings comparison results and we recently rewrote the Portico holdings comparison tool such that we now provide summary information in a more intuitive layout with each comparison. Also, the CRL report identified a concern that Portico is short of archiving a "critical mass" of journal content. Eileen Fenton, the Portico managing director, participated in a recent ALCTS meeting hosted by Martha Brogan, the Chair of the CRL Certification Advisory Panel, at the ALA 2010 Annual Conference. The purpose of this meeting was to discuss the corpus against which any measurement of critical mass should be made.

4. LESSONS, SURPRISES, AND BENEFITS

While the audit entailed a substantial amount of work for Portico, the interactions with the CRL audit team were pleasant, productive, and beneficial. The CRL audit team was extremely thorough and reviewed in great detail all documentation and samples we provided, the Portico website, and the audit and access interfaces. We appreciated their deep interest in learning about the Portico processes. One substantial benefit from this process is simply the opportunity for external review and validation of the approach and processes employed by Portico in pursuit of our preservation work.

Early in the process, Portico decided it was important to ensure that the CRL audit team understood our preservation philosophy, policies, and workflow. This decision to emphasize education and deep understanding had a large impact on the amount of effort required to complete the audit. Rather than collect documentation and forward it to CRL piece meal, we identified existing and missing documentation, collected and wrote documentation, collated it into portfolios, and wrote cover notes to nearly each document. The logistics of this manual process were time consuming. In the end, the process served Portico well.

It is difficult to measure what impact the CRL certification of Portico has made on decisions others make in regard to Portico participation. Portico's certification has been a point of conversation within discussions we are in with the National Library of Medicine in regard to whether or not Portico may be considered an acceptable archive in regards to Medline indexing (currently, the only acceptable archive is PubMed Central). [4]

Portico has benefited in many ways from going through the audit process. We frequently interact with members of the community and respond to requests for information. We have been able readily to share materials collected and documented during the audit process as part of these dialogues. Another benefit arose from the CRL audit team's interest in speaking with a Portico data export partner. As a result we held debriefing conversations with each of our data export partners. These conversations helped us better define the inter-organizational aspects of a data export and ways we can bring Portico's data transformation expertise to questions that might arise during our partners' work with preservation formats and packaging.

It would benefit managers of repositories of all sizes to evaluate their repository against TRAC or perform a self-assessment of risk via DRAMBORA. Whether any individual repository should be audited by a 3rd party, such as CRL, will depend upon the preservation commitments that repository has made, the uniqueness of the content it preserves, and the importance of the content to the repository's designated community (the community served by the repository [3]). Repositories at a smaller scale than Portico and with a more limited community or preservation commitment will, perforce, not have the same level of documentation as Portico. Whether a 3rd party preservation audit is required of any given repository is a decision that must be made by the community served by that repository.

The greatest benefit to Portico was simply the reassurance to Portico and the ITHAKA Board, to the publisher community, to the library community, and to the greater academic community, that the Portico archive was being rigorously examined by an external party. Portico provides auditor privileges to a maximum of four librarians from each participating library and to representatives at each participating publisher (librarian auditors may audit the entire archive and publisher auditors may audit their own content), but it is important to supplement this independent and individual audit activity with a more extensive and systematic approach, as demanded by a TRAC-oriented audit.

To ensure that Portico's certification remains current, Portico will provide CRL with updated documentation every two years and will continue dialogue with CRL on a variety of topics, including what content Portico should target as high priority for preservation. We are looking forward to an ongoing and active dialog with CRL.

[4] http://www.nlm.nih.gov/pubs/factsheets/j_sel_faq.html#a2

5. REFERENCES

[1] The Center for Research Libraries, *"CRL Report on Portico Audit Findings"*, Chicago, Illinois, 2010, http://www.crl.edu/sites/default/files/attachments/pages/CRL%20Report%20on%20Portico%20Audit%202010.pdf (accessed May 4, 2010).

[2] The Center for Research Libraries & OCLC, *"Trustworthy Repositories Audit & Certification: Criteria and Checklist"*, Chicago, Illinois, 2007, http://www.crl.edu/sites/default/files/attachments/pages/trac_0.pdf (accessed May 4, 2010).

[3] Consultative Committee for Space Data Systems (CCSDS), "Reference Model for an Open Archival Information System (OAIS)", Washington, DC, 2002, http://public.ccsds.org/publications/ archive /650x0b1.pdf (accessed July 11, 2010).

[4] Digital Curation Centre & Digital PreservationEurope, *"DCC and DPE Digital Repository Audit Method Based on Risk Assessment,* v1.0", http://www.repositoryaudit.eu/ (accessed May 4, 2010).

[5] Wikipedia contributors, "Audit," *Wikipedia, The Free Encyclopedia,* http://en.wikipedia.org/w/index.php?title=Audit&oldid=359674809 (accessed May 4, 2010).

[6] Wikipedia contributors, "Certification," *Wikipedia, The Free Encyclopedia*, http://en.wikipedia.org/w/index.php?title=Certification&oldid=358904216 (accessed May 4, 2010).

FULL PAPER

MEASURING CONTENT QUALITY IN A PRESERVATION REPOSITORY: HATHITRUST AND LARGE-SCALE BOOK DIGITIZATION

Paul Conway
University of Michigan
School of Information
105 South State Street
Ann Arbor, MI 48109-1285
pconway@umich.edu

ABSTRACT

As mechanisms emerge to certify the trustworthiness of digital preservation repositories, no systematic efforts have been devoted to assessing the quality and usefulness of the preserved content itself. With generous support from the Andrew W. Mellon Foundation, the University of Michigan's School of Information, in close collaboration with the University of Michigan Library and HathiTrust, is developing new methods to measure the visual and textual qualities of books from university libraries digitized by Google, Internet Archive, and others and then deposited for preservation. This paper describes a new approach to measuring quality in large-scale digitization; namely, the absence of error relative to the expected uses of the deposited content. The paper specifies the design of a research project to develop and test statistically valid methods of measuring error. The design includes a model of understanding and recording errors observed through manual inspection of sample volumes, and strategies to validate the outcomes of the research through open evaluation by stakeholders and users. The research project will utilize content deposited in HathiTrust – a large-scale digital preservation repository that presently contains over five million digitized volumes – to develop broadly applicable quality assessment strategies for preservation repositories.

1. INTRODUCTION

The large-scale digitization of books and serials is generating extraordinary collections of intellectual content that are transforming teaching and scholarship at all levels of the educational enterprise. Along with burgeoning interest in the technical, legal, and administrative complexities of large-scale digitization [2], significant questions have risen regarding the quality and fitness for use of digital surrogates produced by third-parties such as Google or the Internet Archive. Until recently, those who built digital repositories also exercised significant control over the creation of digital content, either by specifying digitization best practices [24] or by limiting the range of digital content forms accepted for deposit and long-term maintenance [29]. For an institution and its community of users to trust that individual digital objects created by third parties are accurate, complete, and intact and to know that objects deposited in preservation repositories have the capacity to meet a variety of uses envisioned for them by different stakeholders, repositories must validate the quality and fitness for use of the objects they preserve.

Information quality is an important component of the value proposition that digital preservation repositories offer their stakeholders and users [12]. For well over a decade, the cultural heritage community of libraries, archives, and museums has embraced the need for trustworthy digital repositories with the technical capacity to acquire, manage, and deliver digital content persistently [42]. During the past decade, standards-based mechanisms for building and maintaining repository databases and associated metadata schema have emerged to enable the construction of preservation repositories on a scale appropriate to the preservation challenge at hand [26][19]. Significant progress has been made in establishing the terms and procedures for certifying trustworthiness through independently administered auditing processes [40]. In the new environment of large-scale digitization and third-party content aggregation, however, certification at the repository level alone may be insufficient to provide assurances to stakeholders and end-users on the quality of preserved content. One of the grand challenges of digital preservation is for repositories to establish the capacity to validate the quality of digitized content as "fit-for-use," and in so doing provide additional investment incentives for existing and new stakeholders.

2. LITERATURE REVIEW

Critics of quality: Although large-scale digitization programs have their vocal advocates [13], scholars, librarians, and the preservation community increasingly are raising concerns about the quality and usability of image and full-text products [34]. For example, Bearman [4], Duguid [18], and Darnton [14] cite scanning and post-production errors in early iterations of Google's book digitization program. Tanner [39] finds a high level of error in text conversion of newspapers. S. Cohen [9] suggests that quality issues will arise most strikingly when entire books are printed on demand. Schonfeld [37] concludes that only the full comparison of original journal volumes with their digital surrogates is sufficient before hard copies can be withdrawn from library collections. Attempting to sort through the commentary, D. Cohen [18] identifies a fundamental need for research: "of course Google has some poor scans—as the saying goes, haste makes waste—but I've yet to see a *scientific* survey of the overall percentage of pages that are unreadable or missing (surely a miniscule fraction in my viewing of scores of Victorian books)."

Information quality definitions: The quality of digital information has been a topic of intense research and theoretical scrutiny since at least the mid-1990s. At a definitional level, Garvin [20] identifies five discrete approaches to understanding quality, two of which – product-based and user-based – are particularly relevant to the proposed research. Models for information quality have emerged from important empirical research on data quality [41] and have been adapted for the Internet context [25] Research derived from business auditing principles [6] and information science theory [35] grounds the analysis of information quality in the language of credibility and trust. Research informed by archival theory has also addressed the importance of information quality [43]. Although the emergent models are quite inconsistent in terminology, they provide a comparable theoretical foundation for research on quality in large-scale digitization. The research design described here joins the relatively objective product-based findings on digitization quality with the more subjective evaluation judgments of a user-based approach.

Fitness for use: Stvilia [38] builds on the commonality that exists in information quality models, and focuses special attention on the challenge of measuring the relationship between the attributes of information quality and information use. In adopting the marketing concept of "fitness for use," he recognizes both the technical nature of information quality and the need to contextualize "fitness" in terms of specific uses. Stvilia establishes and tests a useful taxonomy for creating quality metrics and measurement techniques for "intrinsic qualities" (i.e., properties of the objects themselves). In the context of digitization products, intrinsic quality attributes are objectively determined technical properties of the digitized volume, derived from the results of digitization and post-scan image processing. By distinguishing measurable and relatively objective attributes of information objects from the usefulness of those objects, Stvilia establishes a viable research model that can be applied to the measurement of the quality of digitized books within particular use-cases.

Use-cases: Quality judgments are by definition subjective and incomplete. From the perspective of users and stakeholders, information quality is not a fixed property of digital content [11]. Tolerance for error may vary depending upon the expected uses for digitized books and journals. Marshall [31, p. 54] argues that "the repository is far less useful when it's incomplete for whatever task the user has in mind." Baird makes the essential connection between quality measurement and expected uses in articulating the need for research into *goal directed metrics* of document image quality, tied quantitatively to the reliability of downstream processing of the images." [3, p. 2] Certain fundamental, baseline capabilities of digital objects span disciplinary boundaries and can be predicted to be important to nearly all users. Use-cases articulate what stakeholders and users might accomplish if digital content was validated as capable of service-oriented functions [7].

Error measurement: The literature on information quality is relatively silent on how to measure quality attributes of very large collections of digitized books and journals, created as a combination of page images and full-text data by third party vendors. Lin [28] provides an excellent review of the state of digital image analysis (DIA) research within the context of large-scale book digitization projects. Because Lin's framework is determined by ongoing DIA research problems, his "catalog of quality errors," adapted from Doermann [17], may be overly simplistic; but his work is most relevant because it distinguishes errors that take place during digitization [e.g., missing or duplicated pages, poor image quality, poor document source] from those that arise from post-scan data processing [e.g., image segmentation, text recognition errors, and document structure analysis errors]. Lin recognizes that, in the future, quality in large-scale collections of books and journals will depend on the development of fully automated analysis routines. The state of the art in quality assurance today depends in large measure upon manual visual inspection of digitized surrogates or the original book volumes [27]. Although the research design is oriented toward the possibility of eventual automated quality assurance, data gathering will be based fundamentally on manual review of statistically valid samples of digitized volumes.

3. HATHITRUST TEST BED

HathiTrust is a digital preservation repository that was launched in October 2008 by a group of 25 research universities, including the Committee on Institutional Cooperation [the Big Ten universities and the University of Chicago] and the University of California system.[1] At present [July 2010] HathiTrust consists of 6.2 million digitized volumes ingested from multiple digitization sources (primarily Google). HathiTrust is a large-scale exemplar of a preservation repository containing digitized content 1) with intellectual property rights owned by a variety of external entities, 2) created by multiple digitization vendors for access, and 3) deposited and held/preserved collaboratively. HathiTrust is also a technological environment for collaboratively addressing challenges in duplication, collection development, and digital preservation that are common to all libraries. The repository is in the midst of a rigorous certification audit by the Center for Research Libraries using the TRAC [40] framework. HathiTrust is supported by base funding from all of its institutional partners, and its governing body includes top administrators from libraries and information offices at investing institutions [44].

HathiTrust is highly organic, posing interesting challenges for quality assessment, and at the same time making it an ideal test-case for quality research. Large portions of HathiTrust can amount to an information quality "moving target," because the repository overlays existing copies of works digitized by Google with improved versions as Google makes those versions available (between 100,000 and 200,000 volumes are improved and replaced in this way each month, on average). HathiTrust also is growing rapidly, having increased in size by a monthly average of 230,000 volumes in 2009. This volubility challenges the assignment of quality projections across the entire repository. HathiTrust, however, possesses the technical infrastructure and the type of digital content required to develop quality metrics, validate those metrics with users, and assess quality changes over time. The findings of this research will be broadly applicable to the current digital repository environment, ranging from smaller and somewhat stable repositories to large-scale evolving digital preservation services such as HathiTrust.

4. DIGITIZATION QUALITY AND ERROR

The research design is innovative in part for its effort to rethink what quality means within the context of preserved digital content. Until very large-scale digitization forced this issue to the forefront, the preservation community attempted to influence digitization quality through adherence to best practices that the community itself promulgated [24]. Successful implementation of guidelines enables the vertical integration of content creation, content delivery, and content preservation at a scale that seemed large ten years ago but which now pales in comparison to the efforts of third party digitizers such as Google. With vertical integration also comes the possibility of controlling digitization workflows that span the entire conversion-to-preservation process.

Today's digital content environment is marked by distributed responsibility for content creation and a trend toward collaborative responsibility for long-term preservation and access [10]. Increasingly, preservation repositories take what they can get, with, at best, assurances from the publisher/creator that the submitted content meets the original purposes or those deemed appropriate by the creator/publisher [30]. In a distributed content creation environment, it may be both infeasible and inappropriate to validate digitization quality against a community "gold standard."[2] Rather, preservation repositories may have to establish benchmarks that represent the best efforts of the content creator. Such a "bronze standard" recognizes the limitations of large-scale digital conversion and reorients quality assurance toward detecting and remedying errors that may occur at stages of the conversion process.

Within the context of a large-scale preservation repository, our research adapts Stvilia's [38] model of intrinsic quality attributes and Lin's [28] framework of errors in book surrogates derived from digitization and post-scan processing. The error measurement model for the project design recognizes that errors originate from some combination of problems with (a) the source volume (original book), (b) digital conversion processes (scanning and OCR conversion), and (c) post-scan enhancement processing. The research design draws on data from four years of quality review compiled by the University of Michigan Library (MLibrary) as part of the ingest of over five million volumes into HathiTrust. The MLibrary quality review manual, which defines and illustrates eight digitization errors evaluated in books deposited in HathiTrust for the past three years, is available online.[3]

Table 1 presents the distribution of critical level of eight errors identified by University of Michigan library staff over a four-year period. A critical error is one whose presence in one or more of a random sequence of 20 pages is sufficiently severe to render the volume unusable. The table shows the total number of volumes ingested into HathiTrust in a given year, the total number and total percentage of volumes inspected for

[1] HathiTrust. http://www.hathitrust.org/

[2] Federal Agencies Digitization Guidelines Initiative. http://www.digitizationguidelines.gov/

[3] http://www.hathitrust.org/documents/UM-QR-Manual.pdf

Critical Error Type	*Cause*	*May 2006-April 2007*		*May 2007-April 2008*		*May 2008-April 2009*		*May 2009-April 2010*		*TOTAL*
Thick text	scanning	189	0.57%	70	0.19%	19	0.06%	144	0.81%	422
Broken text	scannng	518	1.57%	121	0.33%	76	0.26%	64	0.36%	779
Blurred text	scanning	252	0.76%	40	0.11%	10	0.03%	54	0.30%	356
Obscured text	source	57	0.17%	35	0.09%	21	0.07%	8	0.04%	121
Warpped page	post-scan	47	0.14%	37	0.10%	14	0.05%	22	0.12%	120
Cropped text block	post-scan	424	1.28%	246	0.67%	100	0.34%	67	0.38%	837
Cleaning	post-scan	208	0.63%	214	0.58%	1256	4.23%	439	2.46%	2117
Colorization	post-scan	3250	9.83%	272	0.74%	35	0.12%	19	0.11%	3576
Volumes ingested		288,044		460,620		2,523,049		1,665,167		4,936,880
Volumes reviewed (20 pages/vol.)		33,047		36,981		29,677		17,850		117,555
Ingested/Received		11.47%		8.03%		1.18%		1.07%		2.38%

Table 1. Incidence of critical error in volumes ingested into HathiTrust, 2006-10.

errors using an online logging system built at Michigan. The summary inspection data shows a declining proportion of volumes inspected over time, due to confidence in the inspection process garnered after the first two years of quality assurance work across approximately 70,000 volumes. The table also shows the relatively low rate of critical error and the low absolute number of volumes with critical errors. Errors in post-scan image manipulation (cleaning, colorization, cropping) account for a very large portion of the errors logged. The number of volumes with errors in a given year cannot be totaled, due to the fact that volumes with errors most likely display multiple types of critical error. For example, volumes with warped pages are also likely to have pages with blurred text. The research design adjusts for a flaw in the Michigan model of error inspection, which does not allow for disambiguating error incidence.

The research design builds on the Michigan error detection framework, first by determining the nature and level of intrinsic quality error at three levels of abstraction: (1) data/information; (2) page-image; (3) whole volume as a unit of analysis. Within each level of abstraction exist a number of possible errors that separately or together present a volume that may have limited usefulness for a given user-case scenario. At the data/information level, a volume should be free of errors that inhibit interpretability of text and/or illustrations viewed as data or information on a page. At the page-image level, a volume should be free of errors that inhibit the digital representation of a published page as a whole object. At the whole-volume level, a volume should be free of errors that affect the representation of the digital volume as a surrogate of a book. Errors originate from some combination of problems with the source volume (original book) or digitization (scanning, post-processing).

A major goal of the study is to define meaningful distinctions in severity of error and to validate those distinctions within specific use cases. The project design's error incidence model in Table 2 modifies the Michigan error model (bolded items) by adding reference to possible errors with book illustrations [23], OCR full-text errors, and errors that apply fully to an entire volume. Error detection must account for frequency and severity and be contextualized by level of abstraction. The development of specific judgments of severity of error requires assessment on ordinal scales instead of the binary distinctions between critical and non-critical error utilized presently.

5. RESEARCH MODEL AND METHODOLOGY

The overall design of the research project consists of two overlapping investigative phases. Phase one will define and test a set of error metrics (a system of measurement) for digitized books and journals. Phase two will apply those metrics to produce a set of statistically valid measures regarding the patterns of error (frequency and severity) in multiple samples of volumes drawn from strata of HathiTrust. The design of each phase is anchored by a specific research question that drives the associated data gathering, analysis, and user validation activities.

LEVEL 1: DATA/INFORMATION
1.1 Image: thick [character fill, excessive bolding, indistinguishable characters]
1.2 Image: broken [character breakup, unresolved fonts]
1.3 Full-text: OCR errors per page-image
1.4 Illustration: scanner effects [moiré patterns, halftone gridding, lines]
1.5 Illustration: tone, brightness, contrast
1.6 Illustration: color imbalance, gradient shifts
LEVEL 2: ENTIRE PAGE
2.1 Blur [movement]
2.2 Warp [text alignment, skew]
2.3 Crop [gutter, text block]
2.4 Obscured/cleaned [portions not visible]
2.5 Colorization [text bleed, low text to carrier contrast]
2.6 Full-text: patterns of errors at the page level (e.g., indicative of cropping errors in digitization processing)
LEVEL 3: WHOLE VOLUME
3.1 Order of pages [original source or scanning]
3.2 Missing pages [original source or scanning]
3.3 Duplicate pages [original source or scanning]
3.4 False pages [images not contained in source]
3.6 Full-text: patterns of errors at the volume level (e.g., indicative of OCR failure with non-Roman alphabets)

Table 2. Error incidence model for digitized book and serial volumes.

We refer to "validation" in our research model in two ways that expressly bridge the product-based findings and the user-based approaches to quality. First, validation also refers to the procedures that engage users in identifying the distinctive combination of digitization errors that apply to a given use-case. Second, validation refers to the data analysis routines that demonstrate the statistical power of the error analysis to measure the difference between observed and benchmarked volumes.

Validation through user-based feedback provides a "reality check" that statistically determined findings on quality properly describe the" fitness for use" of digitized volumes.

5.1. Use Case Scenarios

The aim of user-based validation is to confirm that the metrics we have chosen through statistical analysis and then assigned to use cases resonate with users who specify particular use scenarios for HathiTrust content. The development of use-cases is a method used in the design and deployment of software systems to help ensure that the software addresses explicit user needs. Within broad use-cases, individual users can construct stories or scenarios that articulate their requirements for digital content [1]. The research model utilizes use-case design methods to construct specific scenarios for four general purpose use-cases that together could satisfy the vast majority of uses:

Reading Online Images: A digitized volume is 'fit for use' when digital page-images are readable in an online, monitor-based environment. Text must be sufficiently legible to be intelligible [16][32]; visual content of illustrations and graphics are interpretable in the context of the text [23][5], where the envisioned use is legibility of text, interpretability of associated illustrations, and accurate reproduction of graphics sufficient to accomplish a task.

Reading Volumes Printed on Demand: This case refers to printing volumes (whole or substantial parts) derived from digital representations of original volumes upon request [21]. For a volume to be suitable for a print on demand service, it must be accurate, complete, and consistent at the volume level. A print copy is two steps removed from the original source, yet it serves as a ready reference version of the original.

Processing Full Text Data: Most expansively, this use-case specifies the suitability of the underlying full text data for computer-based analysis, summarization, or extraction of full-text textual data associated with any given volume [15]. For a volume to be acceptable for full-text processing, it must support one or more examples of data processing, including image processing and text extraction (OCR), linguistic analysis, automated translation, and other forms of Natural Language Processing [36], most typically applied in the digital humanities.

Managing Collections: This use-case encompasses collaboration among libraries to preserve print materials in a commonly managed space, as well as the management and preservation of the "last, best copy" of regionally determined imprints [33][37]. For digital surrogates to support collection management decision making, digitized volumes must have a sufficiently low frequency or severity of error that they can serve as replacement copies for physical volume.

5.2. Phase One – Metrics

Research Question 1: What is the most reliable system of measurement (metrics) for determining error in digitized book and serial volumes? As a point of departure, the research design hypothesizes a state of image and text quality in which digitized book and serial benchmark-volumes from a given vendor are sufficiently free of error such that these benchmark-surrogates can be used nearly universally within the context of specific use-case scenarios. In the first phase of the research project, we will explore how to specify the gap between benchmarked and digitized volumes in terms of detectable error. The outcome of the first-phase data

gathering and analysis will be a highly reliable, statistically sound, and clearly defined error metrics protocol that can be applied in phase two to measure error-incidence in HathiTrust volumes. Addressing the first research question will require the research team to identify benchmark digitized volumes and create a data model for measuring the presence of error within a given digitized volume.

Identify Benchmark Volumes: The detection and recording of errors will be undertaken in reference to the very best examples of digitized volumes from a given vendor (e.g., Google), rather than in reference to an externally validated conversion standard. Benchmarks are volumes that have no errors that inhibit use in a given use-case. Such "bronze standards" will serve as the basis for developing training materials, establishing the point of departure for coding the severity of error, and validating quality baselines as part of the evaluation strategy.

Draw Samples: A programmer, with the guidance of a statistician, will draw multiple small random samples from selected strata of HathiTrust deposits by manipulating descriptive metadata for individual volumes (e.g., data of publication, LC classification, language). The purpose of sampling is to gather a representative group of volumes to test and refine the error definition model and determine the proper measurement scales for each error, rather than to make projections about error in a given strata population.

Code Errors: Staff and student assistants working in two research libraries [Michigan, Minnesota] will carry out whole-book manual review on the sample volumes, compiling the results initially in a spreadsheet designed by the graduate student research associate. The distinctive data gathering goals are: (1) to determine mechanisms for establishing gradations of severity within a given error-attribute; (2) to establish the threshold of "zero-error" that serves as a foundation for establishing the frequency of error on a given volume-page; and (3) confirm the estimates of error-frequency that determine specifications for the error review system.

Refine Error Data Model: The fundamental units of data in the research design are recorded frequency (counts) and severity (on an ordinal scale) of human-detectable error in either image or full-text data at the page level. The overall data model allows for errors related to image, full-text and illustrations within single pages (e.g., broken text, OCR errors, scanner effects on illustrations), or digitization errors that effect the readability of page images or associated full-text (e.g., blur, excessive cropping), and errors that are counted in pages but applied to entire volumes (e.g., missing or duplicate pages).

Determine Error Co-Occurrence: The research project will test the validity of each error measure in terms of the extent of co-occurrence of pairs of errors. Two measures are completely independent if the two errors never occur together on the same page, whereas two measures are totally dependent if the two errors always occur on the same page. For errors that occur with reasonable frequency, we will test the null hypothesis that error types are independent of each other using a 2 x 2 contingency table and Fisher's exact test for independence. This test for significance is used when the chi square expected frequencies are small. The measure of co-occurrence is a valid way to identify discrete error measures and, possibly, to reduce the number of error measures required to derive an overall measure of quality for a given volume.

5.3. Phase Two – Measurement

Research Question 2: What are the most accurate and efficient measures of error in HathiTrust content, relative to benchmarked digitized volumes? To examine the second question, results based on data analysis for Research Question 1 will be used to create and test measurement strategies for gathering error data from multiple diverse samples of volumes deposited in HathiTrust. Detection of error in digitized content is accomplished through the manual inspection of digital files and sometimes through comparison of digitized volumes with their original sources. The net results of the second phase of the project will be measures of error, aggregated to the volume level, that have as high of a level of statistical confidence as is possible to obtain through manual review procedures. Additionally, the outcome in phase two will be reliable estimates of the distribution of error in the population strata related to the analyzed samples.

Establish Sampling Strategies: The research project will design and implement procedures to draw random samples of volumes for manual inspection and to establish systematic page sampling specifications for review inside any given volume. Data analysis is designed to identify (1) the smallest sample size that can be drawn and analyzed to produce statistically meaningful results; (2) when is it most appropriate to utilize whole-book error analysis as opposed to examining an appropriately sized and identifiable sub-set of page images for a given book; and (3) when is it necessary and appropriate to examine errors in original source volumes as opposed to limiting analysis to digital surrogates. The size and number of volumes and samples depends upon the desired confidence interval (95%) and estimates of the proportion of error within the overall population. Based on three years of error assessment at Michigan, we expect the incidence of any given error to be well below 3%. Given this low probability of error, but where such error may indeed be catastrophic for use, the initial sampling strategy will utilize the medical clinician's "Rule of Three" [22], which specifies that 100 volumes or 100 pages sampled systematically in a typical volume will be sufficient to

detect errors with an expected frequency < .03. Larger sample sizes are required for lower estimates of error.

Gather Data from Multiple Samples: Project staff will create, disseminate, and explain training materials to students and staff coders. A coding manual will contain narrative and visual examples of each error in the protocol, along with detailed instructions for coding error in the quality review system. Trained coders in the two participating academic libraries at the universities of Michigan and Minnesota will record the frequency (error counts) and severity (ordinal scale) of error in images and full-text data at the page level, as appropriate. The sampling strategy (outlined above) will determine the coding and analysis procedures in the two libraries. The data gathering design specifies resources in two research libraries sufficient to review and code approximately 5,000 volumes in samples of 100, 200 or 300 volumes per series. Estimates of review productivity, derived from the planning project supported by the Mellon Foundation, call for one hour of analysis and coding per volume, which will generate approximately 40 data values for each page reviewed in each volume. Data from error assessment activities will be collected in a centralized database at Michigan and subjected to data validation, cleaning, and processing routines by the graduate student research associate.

Assess Extent of Inter-coder Consistency: The research will adapt analytical procedures designed to diagnose and address the challenge of detecting and adjusting for the fact that two human beings will see and record the same information inconsistently. The presence of significant levels of inter-coder inconsistency generates error in the statistical evaluation of the findings of quality review undertaken by multiple reviewers in a distributed review environment. One error review procedure will entail multiple reviewers coding the severity of errors in the same volumes. Collapsing severity to a two-point scale (severe/not) will allow the testing of the null hypothesis that the pairs of reviewers code error severity in the same way, using Cohen's Kappa statistic as a measure of agreement. Similar tests assessing the frequency of errors detected will utilize the Chi Square test of significance. The outcome of these analyses will support improved training of coders and establish the lower threshold of coding consistency in a distributed review environment.

Aggregate from Page to Volume and Evaluate Results: The level of detail in error data at the page level will permit statistically significant aggregation of findings from page to volume. Data gathered at the page level for frequency and severity will be aggregated to the volume level to create coordinate pairs that can be plotted for further analysis. Volume-level error aggregation is the foundation for establishing quality scores for digitized volumes based on the relative number and severity of errors across a mix of error attributes. Error aggregates from assembled from samples of volumes will allow reliable projections regarding the distribution of error in HathiTrust strata. Examples of possible strata subject to analysis include date and place of publication, subject classification, and digitization vendor.

6. CONTRIBUTIONS AND IMPLICATIONS

The research design is a significant contribution to the science of information quality within the context of digital preservation repositories, because the design is grounded in the models and methods pioneered by information quality researchers. The research design and the subsequent research project are innovative in their approach to quality definition and measurement, building specific error metrics appropriate for books and journals digitized at a large-scale. The design is also methodologically advanced through its full integration of (1) tools and procedures for gathering data about quality errors in digitized collections, (2) the rigorous analysis of that data to improve confidence in the measures, and (3) statistically significant conclusions about the nature of error in a large scale repository. Quality review processes conducted across two libraries helps ensure that the research findings may be generalized and not simply refer to one library's digital content. The quality metrics that will be developed in the research project are broadly applicable to collections of digitized books and journals other than those deposited in HathiTrust.

New metrics for defining error in digitized books and journals and new, user validated methods for measuring the quality of deposited volumes could have an immediate impact on the scope of repository quality assessment activities and specific quality assurance routines. Measurements of the quality and usefulness of preserved digital objects will allow digital repository managers to evaluate the effectiveness of the digitization standards and processes employed in producing usable content, and provide guidance on ways to alter digital content to improve the user experience. It will also allow repositories to make decisions about preserving digitized content versus requiring re-digitization (where possible). The ability to perform reliable quality review of digital volumes will also pave the way for certification of volumes as useful for a variety of common purposes (reading, printing, data analysis, etc.). Certification of this kind will increase the impact that digitally preserved volumes have in the broader discussions surrounding the management of print collections, and the interplay between print and digital resources in delivering services to users.

7. ACKNOWEDGEMENTS

Planning support has been provided by the Andrew W. Mellon Foundation. The author thanks HathiTrust

executive director John Wilkin and the staff of the University of Michigan Library for providing data and technical support. The research project design was developed collaboratively by a planning team consisting of Jeremy York and Emily Campbell (MLibrary), Nicole Calderone and Devan Donaldson (School of Information), Sarah Shreeves (University of Illinois), and Robin Dale (Lyrasis).

8. REFERENCES

[1] Alexander I. F. & Maiden N.A.M., eds. (2004). *Scenarios, Stories and Use Cases.* New York: John Wiley.

[2] Bailey, C. W. (2010). "Google Book Search Bibliography." Version 6: 4/12/2010. http://www.digital-scholarship.org/gbsb/gbsb.html

[3] Baird, H. (2004). "Difficult and Urgent Open Problems in Document Image Analysis for Libraries," *Proc. of First International Workshop on Document Image Analysis for Libraries* (DIAL '04), Palo Alto, CA, pp. 25-32.

[4] Bearman, D. (2006). "Jean-Noël Jeanneney's Critique of Google," *D-Lib Magazine* 12 (12). http://www.dlib.org/dlib/december06/bearman/12bearman.html

[5] Biggs, M. (2004). "What Characterizes Pictures and Text?" Literary and Linguistic Computing 19 (3), 265-272.

[6] Bovee, M., Srivastava, R. and Mak, B. (2003). "A Conceptual Framework and Belief-Function Approach to Assessing Overall Information Quality," *International Journal of Intelligent Systems.* 18 (1): 51-74.

[7] Cockburn, A. (2000). *Writing Effective Use Cases.* Boston: Addison-Wesley.

[8] Cohen, D. (2010). "Is Google Good for History?" *Dan Cohen's Digital Humanities Blog*. Posting on 12 Jan. 2010. http://www.dancohen.org/2010/01/07/is-google-good-for-history/

[9] Cohen, S. (2009). "Google to reincarnate digital books as paperbacks," *Library Stuff. Information Today, Inc.*, 17 September, 2009. http://www.librarystuff.net/2009/09/17/google-to-reincarnate-digital-books-as-paperbacks/

[10] Conway, P. (2008). "Modeling the Digital Content Landscape in Universities." *Library Hi Tech* 26 (3): 342-358.

[11] Conway, P. (2009). "The Image and the Expert User." *Proceedings of IS&T's Archiving 2009*, Imaging Science & Technology, Arlington, VA, May 4-7, pp. 142-50.

[12] Conway, P. (2010). "Preservation in the Age of Google: Digitization, Digital Preservation, and Dilemmas," *Library Quarterly* 80 (1): 61-79.

[13] Courant, P. (2006). "Scholarship and Academic Libraries (and their kin) in the World of Google," *First Monday* 11 (August). http://131.193.153.231/www/issues/issue11_8/courant/index.html

[14] Darnton, R. (2009). "Google and the New Digital Future," *The New York Review of Books* 56 (20): http://www.nybooks.com/articles/23518

[15] DeRose, S. et al. (1990). "What is Text, Really?" *Journal of Computing in Higher Education* 1 (2): 3-26.

[16] Dillon, A. (1992). "Reading from paper versus screens: A critical review of the empirical literature." E*rgonomics*, 35(10): 1297–1326.

[17] Doermann, D., Liang, J., and Li, H. (2003). "Progress in Camera-Based Document Image Analysis." *Proc. Seventh International Conference on Document Analysis and Recognition (ICDAR'03),* 3 (6): 606-616.

[18] Duguid, P. (2007). "Inheritance and Loss? A Brief Survey of Google Books," *First Monday* 12 (8).

[19] Gartner, R. (2008). "Metadata for Digital Libraries: State of the Art and Future Directions." JISC *TechWatch Report* TSW0801. Bristol, UK: Joint Information Systems Committee.

[20] Garvin, D. A. (1988). *Managing quality: The strategic and competitive edge*. New York: Free Press.

[21] Hyatt, S. (2002). "Judging a book by its cover: e-books, digitization and print on demand." in Gorman, G.E. (ed.) *The Digital Factor in Library and Information Services*. London: Facet Publishing, 112-132.

[22] Jovanovic, B. D. & Levy, P. S. (1997). "A Look at the Rule of Three." *The American Statistician* 51 (2): 137-139.

[23] Kenney, A.R. et al. (1999). *Illustrated Book Study: Digital Conversion Requirements of Printed Illustrations.* (Report to the Library of Congress Preservation Directorate). Ithaca, N.Y.: Cornell University Library.

[24] Kenney, A.R. & Rieger, O.Y. (2000). *Moving Theory into Practice: Digital Imaging for Libraries and Archives.* Mountain View, CA: Research Libraries Group.

[25] Knight, S. (2008). *User Perceptions of Information Quality in World Wide Web*

Information Retrieval Behaviour. (PhD Dissertation). Perth, Australia: Edith Cowan University.

[26] Lavoie, B. (2004**).** *The Open Archival Information System Reference Model: Introductory Guide.* Digital Preservation Coalition Technology Watch Report 04-01. Dublin, OH: OCLC.

[27] Le Bourgeois, et al. (2004). "Document Images Analysis Solutions for Digital Libraries," *Proceedings of the First International Workshop on Document Image Analysis for Libraries* (DIAL'04), Palo Alto, California, pp. 2-24.

[28] Lin, X. (2006). "Quality Assurance in High Volume Document Digitization: A Survey," *Proceedings of the Second International Conference on Document Image Analysis for Libraries* (DIAL'06), 27-28 April, Lyon, France, pp. 319-326.

[29] Lynch, C. A. (2003). "Institutional Repositories: Essential Infrastructure for Scholarship in the Digital Age," portal*: Libraries and the Academy* 3 (2): 327-336.

[30] Markey, K., Rieh, S. Y., St. Jean, B., Kim, J., and Yakel, E. (2007). *Census of Institutional Repositories in the United States: MIRACLE Project Research Findings*. Washington, DC: Council on Library and Information Resources. http://www.clir.org/pubs/abstract/pub140abst.html

[31] Marshall, C. C. (2003). "Finding the Boundaries of the Library without Walls." In. Bishop, A., et al. (eds.) *Digital Library Use: Social Practice in Design and Evaluation.* Cambridge: MIT Press, pp. 43-64.

[32] O'Hara, K. (1996). "Towards a Typology of Reading Goals. Xerox Technical Report." http://www.xrce.xerox.com/content/download/6681/51479/file/ EPC-1996-107.pdf

[33] Payne, L. (2007). *Library Storage Facilities and the Future of Print Collections in North America.* Online Computer Library Center: Dublin, Ohio. www.oclc.org/programs/publications/reports/2007-01.pdf

[34] Rieger, O. (2008). *Preservation in the Age of Large-Scale Digitization: A White Paper.* Washington, DC: Council on Library and Information Resources.

[35] Rieh, S. (2002). "Judgment of Information Quality and Cognitive Authority in the Web," J*ournal of the American Society for Information Science and Technology* 53 (2): 145-161.

[36] Rockwell, G. (2003). "What it Text Analysis, Really?" *Literary and Linguistic Computing* 18 (2): 209-219.

[37] Schonfeld, R. and Housewright, R. (2009). *What to Withdraw? Print Collections Management in the Wake of Digitization*. New York: Ithaka.

[38] Stvilia, B., et al. (2007). "A Framework for Information Quality Assessment," *Journal of the American Society for Information Science and Technology* 58 (12): 1720-1733.

[39] Tanner, S., Munoz, T., and Ros, P. (2009). "Measuring Mass Text Digitization Quality and Usefulness." *D-Lib Magazine* 15 (July/August): 209. http://www.dlib.org/dlib/july09/munoz/07munoz.html

[40] TRAC. (2007). *Trustworthy Repositories Audit & Certification: Criteria and Checklist.* Center for Research Libraries and OCLC. http://www.crl.edu/sites/default/files/attachments/pages/trac_0.pdf

[41] Wang, R. and Strong, D. (1996). "Beyond Accuracy: What Data Quality Means to Data Consumers," *Journal of Management Information Systems* 12 (4): 5-34.

[42] Waters, D. and Garrett, J. (eds.). (1996). *Preserving Digital Information: Report of the Task Force on Archiving of Digital Information.* Washington, DC: Commission on Preservation and Access.

[43] Yeo, G. (2008). "Concepts of Record (2): Prototypes and Boundary Objects," *American Archivist* 71 (Summer): 118-143.

[44] York, J.J. (2009). "This Library Never Forgets: Preservation, Cooperation, and the Making of HathiTrust Digital Library," *Proc. IS&T Archiving* 2009, Arlington, VA, pp. 5-10.

SHORT PAPER

THE IMPORTANCE OF TRUST IN DISTRIBUTED DIGITAL PRESERVATION: A CASE STUDY FROM THE METAARCHIVE COOPERATIVE

Matt Schultz
Educopia Institute
1230 Peachtree Street, Suite 1900
Atlanta, GA 30309

Emily B. Gore
Clemson University Libraries
CB 343001
Clemson, SC 29634-0001

ABSTRACT

Distributed digital preservation is a maturing and appealing solution to the pressing problem of ensuring the survivability of digital content. Like all other digital preservation efforts, distributed digital preservation solutions must communicate trust to their Designated Communities as they continue to mature. The following paper discusses the importance of establishing this trust, retraces the development of TRAC as a reliable tool for evaluating trustworthy repositories, and details the process of the MetaArchive Cooperative's application of TRAC to its distributed digital preservation solution. This process revealed that the current metrics for gauging trust in digital preservation could be readily applied to distributed solutions with great effect. However, because these metrics often presume a more centralized approach to preservation, the process also revealed the need to apply them carefully and with great thought. To underscore this need, three organizational and technical comparisons are made between the MetaArchive's distributed preservation activities and the more centralized model assumed by TRAC and the OAIS Reference Model. The paper concludes with the question as to whether distributed digital preservation needs to be better defined within existing models such as OAIS or through the creation of a new reference model for distributed digital preservation.

1. INTRODUCTION

Distributed digital preservation is a maturing solution to the pressing problem of ensuring that future generations will have access to digital content of scholarly, cultural, political, and scientific value. As framed in the recently published *A Guide to Distributed Digital Preservation*, "...a growing number of cultural memory organizations have now come to believe that the most effective digital preservation efforts in practice succeed through some strategy for distributing copies of content in secure, distributed locations over time." [13]

Indeed, many projects and service models are actively addressing the need for digital preservation in this geographically distributed fashion. Among these are LOCKSS (Lots of Copies Keep Stuff Safe) and Private LOCKSS Networks (PLNs) such as the MetaArchive Cooperative, ADPNet, PeDALS, and Data-PASS (to name just a few); data grid solutions such as Chronopolis; and cloud-based initiatives such as DuraCloud. These projects and services represent a strong approach that ensures that digital assets can survive well into the future in the face of such threats as natural disasters, human error, and technological obsolescence.

Just like the more centralized institutional or shared repository solutions that have comprised some of the early foundational efforts in the field of digital preservation at large, these distributed digital preservation efforts must focus attention on the issue of communicating trust to their Designated Communities as they continue to mature.

2. IMPORTANCE OF TRUST

Trust is defined as the "reliance on the integrity, strength, ability, and surety of a person or thing." [6] When establishing a preservation service model, especially one with a distributed membership, like the MetaArchive Cooperative and other distributed digital preservation efforts, it is important that trust be at the center. Members need to trust each other, trust the leadership, and trust the preservation system itself. Establishing and maintaining trust can be a daunting task even when colleagues and peers, as opposed to vendors, control, manage and maintain the network. A cooperative model is designed to be much like a democracy, where members take ownership and voice concerns, opinions and shape future directions.

In an interview published in 2000 in *RLG DigiNews*, Kevin Guthrie, then-President of JSTOR, indicates that

establishing trust in 3rd party vendors is "important because the goal is to be able to establish a relationship whereby a library can rely on a third party to provide a service that has been a core function of a library; that is, archiving." [8] The MetaArchive Cooperative supports that belief and arguably enhances it by philosophically and practically striving to enable libraries to work collaboratively to archive their own materials in a trustworthy manner. The MetaArchive Cooperative (www.metaarchive.org) is a community-based network that coordinates low-cost, high-impact distributed digital preservation services among cultural memory organizations, including libraries, research centres, and museums.

Cooperative, distributed digital preservation relationships may be favorable to individual institutions due to both the cost-effectiveness of the approach, which capitalizes on the existing infrastructures of cultural memory organizations rather than requiring the establishment of external services, and the implied sustainability of an alliance of institutions working together. If nothing else, the current economic situation has forced libraries to realize that content in "silo" repositories could be at greater risk as institutional priorities, funding streams, and the greater economy fluctuates. There is greater trust in at least the medium-term sustainability of collaborative efforts than in local efforts where the reduction or elimination of funding for one year can have dire consequences. In collaborative relationships, economic crises at one or two institutions have less of an impact on the collaboration as a whole.

When prospective members consider joining an organization like the MetaArchive Cooperative, trust is arguably the main element they are looking for – they are asking if they can trust the organization, the partners and the technology with the critical assets they are charged to manage for the long-term. In the paper *Creating Trust Relationships for Distributed Digital Preservation*, Walters and McDonald state that, "the concept of trust and its manifestation between institutions as an essential element in designing digital preservation systems − both technical and organizational − is critical and appears in the organizational level needs of the *CRL/RLG-NARA Trustworthy Repositories Audit and Certification (TRAC): Criteria and Checklist*." [7]

3. TRAC

The origin of TRAC itself is in trust relationships and alliances among key organizations. The call for a "network of trusted archives" initially drove the creation of the trusted digital repositories concept as well as influenced the development of the *Reference Model for an Open Archival Information System* (OAIS). [2] As an OAIS-approved follow-on activity, TRAC and the actual metrics development also evolved through these same relationships. The RLG-NARA Task Force on Digital Repository Certification obtained valuable alliances with the then-new Digital Curation Centre, as well as colleagues in Germany directing the *nestor* project. A critical alliance with the Center for Research Libraries (CRL) also emerged. In 2005, the Center for Research Libraries was awarded a grant by the Andrew W. Mellon Foundation to develop the procedures and activities required to audit and certify digital archives. The CRL Certification of Digital Archives Project worked closely together with the RLG-NARA task force to redevelop the audit metrics and provided critical opportunities to develop and test the audit process itself. This practical testing, along with the DCC test audits that led to the development of DRAMBORA, contributed greatly to filling the gaps identified in the earlier draft, *Audit Checklist for the Certification of Trusted Digital Repositories*.

The final version of TRAC was published in February 2007 with 84 criteria broken out into three main sections: Organizational infrastructure; Digital object management; and Technologies, technical infrastructure, and security. It provides tools for the audit, assessment, and potential certification of digital repositories; establishes the documentation requirements for audit; delineates a process for certification; and establishes appropriate methodologies for determining the soundness and sustainability of digital repositories.

It currently serves as a de facto standard for repository audit and is being actively used by organizations as both a planning and self-assessment tool. Additionally, it continues to serve as the basis of further audit and certification work, including the National Science Foundation-funded CRL project, Long-Lived Digital Collections. [5]

4. METAARCHIVE COOPERATIVE SELF AUDIT

A recent effort has detailed for the larger community (including prospective and non-members) the organizational and technological trust foundations of one successful and growing distributed digital preservation solution. Between June and December 2009, the MetaArchive Cooperative worked with an outside evaluator to conduct a self-audit using the *Trusted Repositories Audit & Certification: Criteria & Checklist* (TRAC). [1] The Cooperative makes use of the LOCKSS (www.lockss.org) open source software to dark archive multi-format digital collections. Collections being preserved in the MetaArchive network include electronic theses, digitized photographs and manuscripts, websites, oral histories, and many others. This content is available to the content contributor alone in the event of catastrophic loss of its original content—thus enabling the retention and preservation of the many important works that cannot be openly shared at this time due to intellectual property and other concerns.

4.1. Self Audit Results

The results of the MetaArchive Cooperative's self-audit revealed that the MetaArchive conformed to and addressed the concerns of each of the 84 criteria specified within TRAC. As importantly the assessment helped to identify and prioritize at least 15 activities to be reviewed and/or enhanced over the course of 2010 and 2011. [12] The success of this process made it clear that current metrics for gauging trust in digital preservation could be readily applied to distributed solutions, it also underscored the need to apply them carefully and with great thought.

5. DISTRIBUTED SELF AUDIT METHODS

Assessing the MetaArchive Cooperative revealed that an evaluator at work in this distributed digital preservation environment must be willing to invest a fair amount of time engaging with repository staff through a careful and synthesized analysis in at least three ways:

- The first of these involves systematically coming to grips with the design solution of the repository. This can be done through extensive reading of internal and published documentation and conducting multiple interviews with repository staff. Specifically, an evaluator must ask questions regarding how the repository is organized to effect preservation, and how the underlying technology both facilitates and constrains that organization appropriately.
- The second area of analysis involves comparing and contrasting this overview of the repository with the OAIS Reference Model, and its functional recommendations for building a trustworthy repository.
- Finally, the evaluator must grapple with the concerns embedded in TRAC itself, and ensure that in pursuing the objective of applying OAIS frameworks and definitions to a repository's activities, the evaluation fairly accomplishes its core goal: that of gauging genuine degrees of trust and best practice within the repository.

Though OAIS seeks to apply its functional elements in responsible ways to diffuse models such as those of federated repository endeavours, a centralized model for preservation is largely at focus in OAIS and TRAC. [4] This is no doubt because most digital preservation initiatives, even those, such as the Hathi Trust (http://www.hathitrust.org) that have pursued trustworthy federated approaches have tended to situate each of the OAIS functional elements and roles within single repository spaces for various administrative and technical reasons. For reasons of this precedent an evaluator of a distributed digital preservation network may be required to extrapolate out some of the OAIS Reference Model's elements when necessary and look for their representation across diffuse locations and multiple roles.

5.1. Drawing Fair Comparisons

Three examples that demonstrate the need for such extrapolations stand out from the MetaArchive Cooperative's self-audit.

- *Central vs. Distributed Infrastructure*: this first example sheds light on the importance of being able to draw some proper distinctions between a distributed digital preservation effort's network server environment and its web-like representation of a "repository", in contrast to the more unified and centrally housed infrastructure that tends to be standard to many other digital preservation solutions.
- *Push vs. Pull on Ingest*: this second example highlights the behaviour of the LOCKSS software and its "pull" scheme of ingesting submission information packages (SIPs), and constructively contrasting this with the typical "push" scheme facilitated by many repositories (electronic ETD submissions for institutional repositories as one example).
- *Dark Archiving & Designated Communities*: the third example involves properly addressing the OAIS Reference Model's notions of Access and Designated Communities (Producers/Consumers) in light of the MetaArchive's dark archive approach to bit-preservation and the format agnostic designations of LOCKSS.

5.1.1. Central vs. Distributed Infrastructure

Though the OAIS Reference Model and TRAC both acknowledge that there are multiple ways to organize a repository's infrastructure, the documents themselves overwhelmingly have related a more centralized approach to designing and operating a digital preservation solution. The MetaArchive Cooperative (along with other PLNs, Chronopolis, and other initiatives) has established a distributed network of linked servers that cooperate to mutually store, manage and refresh contributed content at the bit-level. This methodology holds that replications of content that are geographically distributed and maintained on multiple servers in highly secure networks stand the greatest chance of meeting the integrity and longevity standards that the cultural memory field must strive to achieve.

During the course of researching the organizational and functional design of the Cooperative for self-auditing purposes, it became clear that the conceptions of the more stationary and routine operations of a traditional archival "repository" in TRAC had to be mapped to an understanding of the more dynamic and automated changes of state that are inherent to the software operations of LOCKSS. Clarifying this

distinction allowed for a proper response to a central concern within OAIS and TRAC: the fixity or integrity of the content.

LOCKSS, for example, engages in a vigilant, and automated process of verifying that the geographically dispersed copies that have been ingested from a content contributor's source are consistent with that source and with one another. It handles this through the use of a voting and polling scheme between the linked servers with mutual copies of content, and relies on temporary checksum comparisons. Indeed, LOCKSS distinguishes itself from perhaps more static repositories by actively anticipating the potential for corruptibility and has developed a recovery scheme in the face of such eventuality by first of all refusing to rely on long-term validation through the maintenance of checksums – which are themselves easily corruptible. [10] Rather it leverages the validation power of a network of redundant servers, and maintains an open re-ingest stream to the authoritative source, once corruption of a copy is detected.

This is quite different than running digests on a single copy of an ingested digital object as it resides or is migrated on disk/tape and then comparing its hash value to a previously generated checksum, which requires its own set of long-term curatorial data management. This, latter scheme is encouraged by OAIS and TRAC in its prescriptions for content fixity, and is implemented and relied upon by many centralized repositories. Though the concern is one for the content's integrity, in and of itself, this approach often only alerts to the occurrence of file corruption, rather than going beyond this to trigger an automated assist in its diagnosis or recovery.

As an evaluator applying TRAC to the Cooperative, while at the same time trying to genuinely address the concern for the content's integrity that resides around this issue of fixity, it became clear through this careful comparison that the emphasis for this LOCKSS-based network needed to be directed differently. The emphasis needed to be placed less on managing and reporting on the veracity of the fixity data itself (though not unimportant), and more so on being able to report on the rate and nature of content repair and re-ingest, so that any disruptions to network activity could be more properly diagnosed and mitigated. To this end the central staff and membership of the MetaArchive Cooperative have begun experimenting with the rich information handling of the LOCKSS daemon in order to provide timely and actionable reports on the status of the network's operations. Progress on this front is being accomplished with great effect through integrations between the LOCKSS daemon and in-house data reporting tools developed by MetaArchive.

5.1.2. Push vs. Pull on Ingest

In many centralized repositories a content contributor is provided a submission pathway whereby they are charged with handing their digital object(s) off to repository specialists. This hand-off typically occurs in a format that can be easily managed or migrated by the repository for the sake of long-term preservation. Occasionally this places the content contributor in front of an access interface that will accept various user-generated metadata concerning the digital object(s), and a mechanism for uploading these objects, as Submission Information Packages (SIPs). At that point the repository takes over and shepherds the digital object(s) through a series of processes to prepare the objects for long-term storage, management, and dissemination. The pathway is thus a process of "pushing" content into an archive, which aligns quite comfortably with our unquestioned protocols for donating artefacts to traditional archives. It is also the process most visibly detailed within the OAIS Reference Model [3]—and even more so, in the cultural memory community's use and discussion of this model.

Distributed digital preservation solutions have often taken a "pull" approach that differs somewhat from this paradigm. The MetaArchive Cooperative (via LOCKSS and its web-crawl based ingest mechanism), and Chronopolis (via the use of "holey" BagIt files as one of several ingest mechanisms) are both examples of repositories that can be said to be using a "pull" scheme for obtaining digital objects.

Specifically for the MetaArchive this has meant that central repository staff must work in a coordinated fashion with content contributors to ensure that they have prepared their content in structured ways (referred to as 'data wrangling') to ensure a successful and on-going "pull" of their content into the preservation network. Once the content has been prepared this "pull" process is finalized by having a content contributor construct an XML plugin that enforces any inclusion/exclusion rules necessary to identify collection files as they reside on an active web server directory. This plugin is then used by the LOCKSS software to guide a web crawl and perform a harvest of the collection.

An evaluator applying the OAIS Reference Model and TRAC to this arrangement has to recognize and account for the way that various functional elements that would typically be reserved only for repository staff operating under a "push" system, namely the preparing of a SIP to become an Archival Information Package (AIP), need to be looked for in various ways on the side of the content contributors within a "pull" environment. This is because the content contributors take responsibility for preparing their own content for its ultimate preservation state by engaging in the "data wrangling" and defining of their collections for harvest. In the MetaArchive context, this has led to the development of documentation that more explicitly describes the MetaArchive network's expectations regarding content organization and the ingest procedures

that contributors follow. This documentation is thus working to better define the functional point at which a SIP becomes an AIP, and the roles on both sides of the Cooperative community that bear the responsiblility for such transformations.

5.1.3. Dark Archiving & Designated Communities

Though the majority of digital preservation initiatives have linked the priorities of preservation and access quite closely, as in the case of institutional repositories, there are several examples of use cases that make immediate access to preserved materials a secondary priority. Dark archiving, which involves preserving materials for future use with no direct means of access from the repository, is an approach that has been attractive to those with content that needs to be preserved but that is not immediately or openly available for access. This has multiple permutations.

In the case of CLOCKSS (http://www.clockss.org/clockss/Home), publishers and libraries agree that a publisher should retain the authority to provide access to their electronic publications, but that libraries can assume this role under certain conditions. This requires that libraries preserve a copy and restrict access until such a defined "trigger event" has occurred – loss of a publisher or a title no longer being offered for example. Through the use of proxy mechanisms, the *end-user* of a journal's Designated Community may not even notice that the publisher's hosting has switched to that of the library because LOCKSS caches at a library site collect and preserve the original journal content exactly as it was served from the publisher. The switch in many cases appears seamless.

The MetaArchive Cooperative has found the use of the LOCKSS software to be similarly useful for the dark archiving and bit-level preservation of their members' digital collections. As mentioned, a member can construct an XML plugin that enforces any inclusion/exclusion rules necessary to crawl collection files as they reside on an active web server directory. This plugin can then be used by the LOCKSS software to guide a web crawl, perform a harvest, and dark archive the collection on a separate, geographically dispersed server. For such members a copy is thus preserved in the event that the originating web server is unable to provide access to a content contributor for their own institutional purposes.

In the case of the MetaArchive, however, on-going and immediate access for a member's *end-user* Designated Community need not be the ultimate guiding priority. The MetaArchive has taken the de-prioritization of access a step further by avoiding the requirement that members select collections that are "dissemination worthy," or that lend themselves to any foreseeable exhibition and use. In fact, members have broad rights of selection when seeking to preserve their collections in the MetaArchive network. Not only is LOCKSS well designed for preserving content in reserve for future *end-user* access scenarios, but it is also format agnostic. This means that members can not only preserve normalized and derivative files that lend themselves nicely to our current notions of future 'readability' and 'understandabilty', but the original bit stream data, and even high quality master files that can be used for any future, as yet unknown, migration or emulation requirements. Under these terms the MetaArchive Cooperative has empowered its members to assume the curatorial responsibility for the decision-making surrounding the preservation of their collections, rather than requiring them to contribute only highly vetted, access-oriented collections in formats that are considered "manageable" by the repository.

A MetaArchive member enters into agreement with other members to mutually preserve one another's collections to guard against the all to too real threats of natural disaster, human error, and technological obsolescence. These are the "trigger events", and when they occur, a member may recover their collection intact from the network, where it has been both technically and legally shielded from any dissemination chain (including to those institutions that hold replicated copies of the content for preservation purposes). Under these terms, a MetaArchive member is the *end-user* for all intents and purposes, and is in a sense both a Producer and a Consumer in OAIS Reference Model terms.

When assessing such repository arrangements with auditing tools like TRAC it is vital that an evaluator be able to de-couple the notion of a Designated Community of Producers and Consumers from the OAIS Reference Model's emphasis on access and use. Though MetaArchive members may not hypothetically choose to preserve files and formats that satisfy our current notions of maintaining future 'readability' and 'understandability', they have been provided a preservation solution that grants them the flexibility to engage their collections on terms that are appropriate to their institutional priorities – which cannot be underestimated in a time when many cultural memory organizations find themselves contending with short-term limited resources but a desire to avoid outsourcing to multiple third party services, in the hopes that they can gradually build expertise and capacity in preservation.

Nevertheless, the concern with useable formats is a natural one, for which LOCKSS has sought to engage for the possible, but by no means impending, approach of widespread obsolescence. [9] [11] Nor is the MetaArchive opposed to monitoring the current and foreseeable usability of its members' collections. The Cooperative's members and central staff remain open to the potential long-term usefulness of the Unified Digital Formats Registry, and if called upon by its members, to exploring integrations with JHOVE2 and DROID,

especially as tools that could enable the MetaArchive to communicate broadly to its membership the number and types of formats being preserved in the network, thereby further empowering them with the information they might need to effect preservation and access for their own Designated Communities as they define them.

6. CONCLUSION

In much the same way that centralized repositories have worked assiduously to prioritize trust as a guiding principle for design and management of their preservation solutions, the maturing field of distributed digital preservation must also communicate the trust relationships that are foundational for a responsible network. When using current tools to accomplish this aim--such as OAIS, TRAC, and successor tools such as the *Metrics for Digital Repository Audit & Certification* being prepared for ISO standardization–distributed digital preservation solutions must make clear the ways that they differ, both organizationally and technically, from more centralized solutions.

The MetaArchive Cooperative has started this process by engaging in a self-audit with these existing tools. The MetaArchive's ability to actively conform to and address the concerns of each of the 84 criteria within TRAC successfully and to use this audit tool to help it schedule 15 items for review and enhancement, demonstrate that TRAC can be a valuable tool for distributed solutions. However, it is important for evaluators to engage in a careful and synthesized analysis of the repository, the standards, and the audit metrics in order to sincerely address concerns and identify new implementations that are compatible with the distinctive activities that are unique to this growing set of distributed preservation endeavours.

It is also worth questioning whether distributed digital preservation needs to be better defined by its community of practice. Abstracted principles that enable discussion, foster understanding, and provide a foundation for assessment are necessary elements in our growing digital preservation arena. It may be time to explore the efficacy of either better defining a distributed digital preservation network within the existing OAIS framework or creating a reference model that explicitly addresses the technological and organizational issues that arise in the distributed preservation network context.

7. REFERENCES

[1] Center for Research Libraries; Online Computer Library Center. *Trusted Repositories Audit & Certification: Criteria & Checklist Version 1.0*. Center for Research Libraries, Chicago, IL, 2007, Available at: http://www.crl.edu/sites/default/files/attachments/pages/trac_0.pdf

[2] Consultative Committee for Space Data Systems. *Reference Model for an Open Archival Information System (OAIS): ISO 14721:*2003, CCSDS Secretariat, Washington, D.C., 2002. Available at: http://www.iso.org/iso/catalogue_detail.htm?csnumber=24683

[3] Consultative Committee for Space Data Systems. *Reference Model for an Open Archival Information System (OAIS): Pink Book, August 2009*, CCSDS Secretariat, Washington, D.C., 2009, pg. 4-49 – 4-51.Available at http://public.ccsds.org/sites/cwe/rids/Lists/CCSDS%206500P11/Attachments/650x0p11.pdf

[4] Consultative Committee for Space Data Systems. *Reference Model for an Open Archival Information System (OAIS): Pink Book, August 2009*, CCSDS Secretariat, Washington, D.C., 2009, pgs. 6-4 - 6-9. Available at:http://public.ccsds.org/sites/cwe/rids/Lists/CCSDS%206500P11/Attachments/650x0p11.pdf

[5] Dale, Robin L. and Emily B. Gore. *Forthcoming* – "Process Models and the Development of Trustworthy Digital Repositories". Information Standards Quarterly, Spring 2010

[6] Dictionary.com. 2010. Definition of *trust*. Accessed on 1st May, 2010. Available at: http://dictionary.reference.com/browse/trust

[7] McDonald, Robert H. and Tyler O. Walters. "Restoring Trust Relationships within the Framework of Collaborative Digital Preservation Federations," *Journal of Digital Information*, Vol. 11, No. 1, 2010

[8] Research Libraries Group. "Developing a Digital Preservation Strategy for JSTOR, an interview with Kevin Guthrie," *RLG DigiNews* 4, no. 4 (August 15, 2000. Available at: http://worldcat.org:80/arcviewer/1/OCC/2007/08/08/0000070519/viewer/file476.html#feature1

[9] Rosenthal, David. "Format Obsolescence: Assessing the Threats and the Defenses", Library Hi-Tech, Vol. 28, Issue 2, 2010, pgs. 195-200. Available at: http://www.emeraldinsight.com/journals.htm?issn=0737-8831&volume=28&issue=2

[10] Rosenthal, David; Robertson, Thomas; Lipkis, Tom; Reich, Vicky; Morabito, Seth. "Requirements for Digital Preservation Systems: A Bottom-Up Approach", *D-Lib Magazine*, Vol. 11, No. 11, 2005. Available at: http://www.dlib.org/dlib/november05/rosenthal/11rosenthal.html

[11] Rosenthal, David; Robertson, Thomas; Lipkis, Tom; Morabito, Seth. "Transparent Format Migration of Preserved Web Content", *D-Lib Magazine*, Vol. 11, No. 1, 2005. Available at: http://www.dlib.org/dlib/january05/rosenthal/01rosenthal.html

[12] Schultz, Matt. "MetaArchive Cooperative TRAC Audit Checklist", Atlanta, GA, 2010. Available at: http://www.metaarchive.org/sites/default/files/MetaArchive_TRAC_Checklist.pdf.

[13] Skinner, Katherine and Matt Schultz Eds. *A Guide to Distributed Digital Preservation*, Educopia Institute, Atlanta, GA, 2010, pg. 6. Available at: http://www.metaarchive.org/GDDP

LEGAL ASPECTS OF EMULATION

Jeffrey van der Hoeven
Koninklijke Bibliotheek
Prins Willem-Alexanderhof 5
2509 LK The Hague
The Netherlands

Sophie Sepetjan
Bibliothèque nationale de France
Quai François Mauriac
75706 Paris Cedex 13
France

Marcus Dindorf
Deutsche Nationalbibliothek
Adickesallee 1
D-60322 Frankfurt am Main
Germany

ABSTRACT

Apart from technical, organisational and economical challenges of long-term access to digital objects, legal implications arise as well. The KEEP project, co-funded by the European Commission under the Seventh Framework Programme (FP7), researched the legal aspects of emulation as part of their research to create a preservation strategy based on emulation. It addresses the legal implications in France, Germany and The Netherlands as well as on European level. Insights have been gained into the legality of transferring content from old media carriers to newer ones and the reproduction of hardware into software emulators. This paper presents the results of the study, conducted by the Bibliothèque nationale de France (BnF), Deutsche Nationalbibliothek (DNB) and the Koninklijke Bibliotheek (KB, the National Library of the Netherlands).

1. INTRODUCTION

Long-term preservation of digital objects not only implies looking after their conservation, but also necessitates the development and execution of strategies to ensure these objects remain accessible and understandable in the future. Apart from the technical challenges this touches upon some legal implications which have been investigated within the KEEP project.

KEEP is a research project co-funded by the European Union under the Seventh Framework Programme (FP7) and stands for *Keeping Emulation Environments Portable* [14]. The project extends on previous work done on emulation such as the Dioscuri project [12] that developed the Dioscuri emulator and the Planets project [17] which amongst others created emulation and migration services. Furthermore, KEEP follows on the recommendations given by the Emulation Expert Meeting held in The Hague in 2006 [10] which stated that emulation is a vital strategy for permanent access but it requires several next steps to become mature. KEEP aims to deliver a strategy that gives permanent access to multimedia content (such as computer applications and console games), not only now but also over the long term. Therefore, it does research into media transfer, emulation and portability of software. In addition to this research a prototype will be developed that can capture data from old physical carriers and render it via emulation. To avoid having the prototype itself becoming obsolete a virtual layer is created that guarantees portability to any computer environment.

2. EMULATION AS A PRESERVATION STRATEGY

Emulation is a proven technology that can be used to cope with obsolescence of hardware and software. It is a technique that supports the recreation of a computer environment (target) on top of another computer environment (host) [13]. Such adaptation is done by an emulator (often a software application but it could also be embedded in hardware). The emulator mimics the functionality of hardware or software, depending on the kind of emulation level is chosen. Each level of a computer environment can be emulated, that is: hardware, operating system or application. The KEEP project focuses on the level of hardware which entails creating virtual representations of real hardware such as a CPU, memory and graphics card. Altogether it forms a virtual computer that is capable of executing native software (e.g. operating system, drivers, applications). Hardware is often well specified and documented in comparison with other levels (e.g. a closed-source operating system such as Microsoft Windows is very hard to emulate accurately). Moreover, an almost unlimited list of emulators and virtualisation software exist mimicking hardware going back as far as the IBM CP-40 in 1966 [20]. This makes it easier to understand the inner workings of hardware components as this software can be examined and re-used.

In the context of preservation, emulation is an attractive solution. By rendering a digital object with an emulator and original software an authentic recreation of that object in its native computer environment can be given, such as WordPerfect 5.1 on MS DOS 5.0 (see figure 1). The advantage of such a strategy is that no

Work presented in this paper is partially supported by European Community under the Information Society Technologies (IST) Programme of the 7th FP for RTD - Project IST-231954. The authors are solely responsible for the content of this paper. It does not represent the opinion of the European Community, and the European Community is not responsible for any use that might be made of data appearing therein.

changes to the digital object is required which offers better conditions to its authenticity. Apart from a computer museum, in some cases emulation is even the only possible way to gain long-term access to digital information as migration does not work for complex digital objects such as software applications (e.g. games), websites or visualisations of data sets.

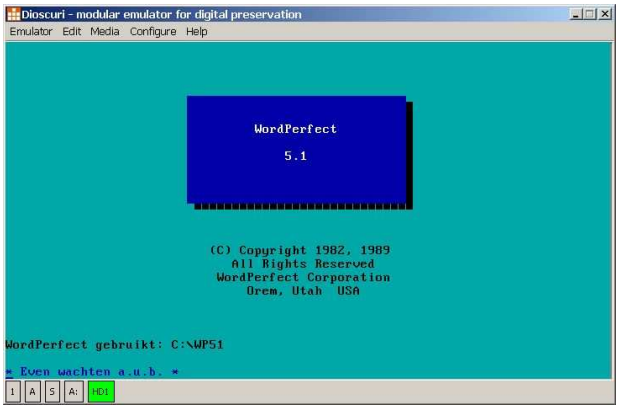


Figure 1. emulator Dioscuri rendering WordPerfect 5.1 on MS DOS 5.0

3. EMULATION ACCESS PLATFORM

KEEP extends on the idea of applying emulation as long-term access strategy in an organisation (e.g. library, museum, company) for its digital collection. To move the emulation strategy from the arena of theoretical discussions into the field of practical solutions some implications have to be solved first. During the first year of the project, the BnF, DNB and KB conducted a survey amongst users of their library. They were asked about their current practices, preferences and desires regarding access to digital information. In total, 644 people responded of which 588 completed the survey. One of the outcomes regarding emulation (figure 2) was that more than half of the respondents (285) noted to have experienced problems accessing old computer files or programs. The technical reasons mentioned were that their current computer could not operate with the old digital file or program (31%). Even so, appropriate media drives seem to be missing (29%) or the media carrier was damaged (17%). Lack of original software is also a significant issue (15%).

This insight is supported by a recent study conducted by the European project PARSE.Insight [16]. Focusing on researchers, 1,209 of almost 1,400 the responding researchers stated that "lack of sustainable hardware, software or support of computer environments may make the information inaccessible" is the most important threat to digital information. However, 81% of the same researchers still preserve their own research data on their local computer.

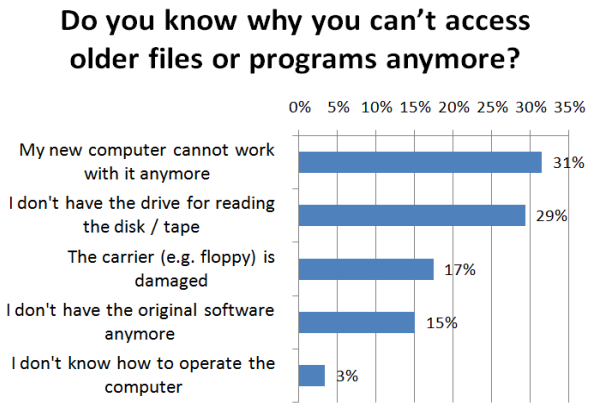


Figure 2. Results on question: do you know why you can't access older files or programs anymore? n = 285 (multiple answers allowed)

Based on this input and on experience with emulation strategies, the following issues require attention:

- Data often resides on obsolete data carriers;
- Original software is required;
- Installing an old computer environment is difficult if not impossible;
- Insufficient descriptive and technical information is available (e.g. manuals, tutorials and other supporting documents).

The KEEP project recognises these technical issues and has envisioned a solution that should help organisations to adopt emulation within their business. The solution is called the Emulation Access Platform (EAP) and will support organisations to:

- Migrate data from old carriers onto newer media carriers;
- Access digital objects in its authentic computer environment using emulators;
- Keep track of sufficient contextual information regarding object and its environment;
- Become independent from current and future computer platforms.

The EAP will consist of three components: Transfer Tools Framework (TTF), Emulation Framework (EF) and the KEEP Virtual Machine (KEEP VM). Figure 3 gives an overview of these components and how they interact with each other and the environment.

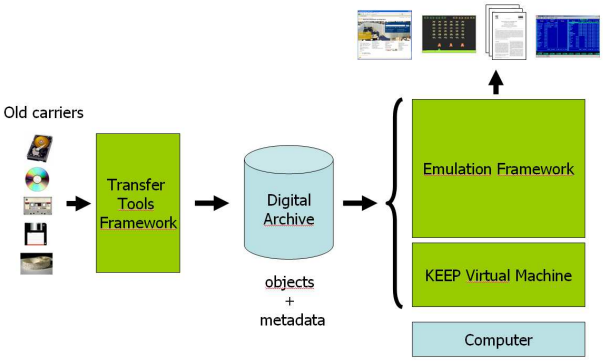


Figure 3. Emulation Access Platform

The first step is to capture the bit stream from old media carriers. This will be done by the TTF integrating already available or new tools to extract raw data from physical media. The data stream (raw bit stream) will then be captured in a container format (disk image) enriched with metadata. After that, the newly created container and metadata can be ingested in the organisation's digital archive via the normal ingest procedures in terms of the OAIS reference model (ISO 14721:2003) [15].

At retrieval time, the object and metadata are disseminated from the digital archive and handed over to the EF. The EF identifies and characterises the digital object using external services such as DROID, PRONOM or another service (e.g. Planets suite [18]). The connection to these services can be customized depending on which services are available and denoted as trustworthy by the organisation's policy. Based on the object's characteristics the emulation environment is constructed automatically. This consists of the appropriate operating system, application software and drivers together with an emulator capable of accurately rendering the object and environment. When preparation is done the EF facilitates a Graphical User Interface (GUI) to the user showing the rendered object in its authentic environment. Additional services such as copying text, making screenshots or recording a video of the rendering process will be supported.

Currently, several prototypes of the EF have been created already based on a set of requirements and a design [19]. The KEEP VM has been specified as well [22] and the requirements for the TTF are almost finalised.

4. LEGAL CHALLENGES

Although emulation is denoted as technically challenging, it has become a more accepted strategy over recent years. The big advantage of not having to migrate all digital objects over time (periodically) saves storage space, time, money and effort and therefore has made this strategy an attractive alternative to migration. However, apart from this technical and economical perspective, the legal conditions should be researched as well [21].

Within the KEEP project a study has been carried out to research the legality of various aspects of emulation [3]. The legal departments of the Bibliothèque nationale de France (BnF), Deutsche Nationalbibliothek (DNB) and the National Library of the Netherlands (KB) worked together with the international law firm Bird&Bird to research the legislation within their own countries as well as European regulations. The legal teams at the libraries are experts in copyright and privacy and find KEEP's research a welcome addition to their journey for better legislation regarding long-term access to cultural and scientific information. The study has been conducted from February 2009 until March 2010 and covers two main topics which are explained in detail in the following sections.

4.1. Media Transfer

To ensure that a digital object will last longer than its media carrier, it has to be transferred to subsequent carriers over time. This process raises some legal issues as reproduction of content is restricted by law. Moreover, various protection mechanisms have been put in place by vendors to prevent users to copy the information stored on the original data carrier. Matters get even more complicated when manufacturers have stopped their businesses or gone bankrupt, leaving their products as 'orphan works' or abandon ware [2]. This leads to a very challenging situation within cultural heritage. On the one hand memory institutions are given the responsibility to preserve the cultural heritage which includes increasing amounts of digital media carriers. On the other hand, most of the digital carriers received are protected against copying and require special treatment to sustain access to the objects contained therein while the legal framework seems very restrictive.

4.2. Emulation of Software and Hardware

For emulation purposes, hardware and their embedded software or semi-conductor products (e.g. chip masks), have to be mimicked. Their inner workings can be understood by reading technical manuals, but in some cases this is not sufficient. Reverse engineering of hardware and software could then be the only way, for example by performing specific tests on original hardware using an oscilloscope or decompiling software.

The issues raised here are whether or not reverse-engineering is lawful where software, hardware or semiconductors are concerned. In addition, as certain hardware may have already been emulated by proprietary or Open Source Software it is therefore worth assessing to which extent these existing emulators could be used by KEEP.

5. CONTEXT & DIRECTIVES

To find out if transferring digital information and using emulation for rendering digital objects is legally possible, the appropriate European and national regulations have to be investigated. The legal study uses the container term 'Multimedia Works' to cover all possible types of content, such as audiovisual content, software and database elements along with off-the-shelf software programs considered on a standalone basis. The legal qualifications of these digital objects differ: they may qualify as computer programs, audiovisual works and/or databases. Moreover, video games are sometimes treated as a special legal category, within national law, as well.

Examining the set of rules and regulations defined by the European Union (also known as the *Community*

Framework) learns that various directives are involved that cover (parts of) the digital objects concerned:

- Directive 2001/29/EC of 22 May 2001 on the harmonization of certain aspects of copyright and related rights in the information society, known as the *Information Society Directive* [8];
- Directive 2009/24/EC of 23 April 2009 on the legal protection of computer programs, known as the *Computer Programs Directive* [9] (replacing the older Directive 91/250/EEC of 14 May 1991) [6];
- Directive 96/9/EC of 11 March 1996 on the legal protection of databases, known as the *Database Directive* [7].

These directives have been researched together with the national laws applicable in France, Germany and the Netherlands.

6. LEGAL IMPACT ON MEDIA TRANSFER

In case of transferring data from media carriers, the study identified various areas of impact outlined in the following sections.

6.1. Intellectual Property Rights

Following the *Information Society Directive* from the EC Multimedia Works are protected by intellectual property rights. This means that reproduction and representation of a protected work must be authorized. Intellectual property rights apply to the work itself rather than to the physical storage media. Therefore, the rules regarding copyright protect the content, whatever the physical medium may be (e.g. floppy disk, optical disk, cartridge). The protection lasts seventy years after the author's death (when the publication is done by an individual) or seventy years after publishing (when the publication is done by a company, or (in the Netherlands) when an employee created the work in the service of an employer).

A special exception exists in the three countries covered by this research (France, Germany and The Netherlands). This exception authorizes reproduction and representation of protected works by institutions responsible for legal deposit (e.g. national libraries), or cultural heritage institutions in general (in the Netherlands no legal obligation exists for preservation publications). This allows them to take appropriate actions such as format migration or media refreshment (transfer of content) to ensure that the digital object in question will not be lost over time.

6.2. Copy Protection Techniques

To protect the duplication of multimedia works and computer programs publishers often use technical measures of protection (TMP). In France, the law dictates that Multimedia Works must be deposited at the BnF with appropriate access codes (software keys) [11]. In Germany, circumvention of TMP is prohibited by the German Copyright Act. However, according to the Code of the German National Library legal deposits shall be done without TMP. If not, the access codes should be given or the TMP must be removed. This applies to multimedia works except for games which are excluded from the legal deposit requirement. Circumvention of technical protection measures is not allowed for these computer programs.

In the Netherlands, technological protection measures on multimedia works and computer programs prevail over the exceptions, unless this is remedied by secondary legislation [4]. As yet, such legislation has not been issued and as a consequence, circumvention of these protection measures is legally not possible.

When circumvention of TMP by the legal deposit institutions is permitted, they are only allowed to do so within their premises. Circumvention by private companies is possible only when they act for these institutions as service providers.

7. LEGAL IMPACT ON EMULATION SOFTWARE

The lawfulness of decompiling computer program environments consisting of operating systems, firmware (e.g. BIOS) and applications should be assessed in line with Article 6 of the *Computer Programs Directive*. This article states that reproduction and translation of source code for the purposes of decompiling are not subject to prior authorisation, provided that (i) these acts are intended to create interoperability between an 'independently created program' and other programs; (ii) these actions are performed by a licensee or lawful user; (iii) the necessary information to obtain such interoperability is not quickly and easily accessible, and (iv) the acts are limited to the portions of the code required for the aim pursued.

Following the line of thought that cultural heritage institutions are lawful users and that they try to create interoperability between old Multimedia Works and current computer environments while no other information is easily at hand, they seem to meet all the predefined conditions. Therefore, decompilation of certain parts of software code is allowed. It would therefore not be necessary to obtain permission in advance from the rights holder.

That said, these conclusions are subject to the development of the emulation platform not requiring the reproduction of a substantial quantity of code from the decompiled computer program, in which case the offence of copyright infringement could be incurred. Likewise, decompiling merely for the purpose of research and analysis without attempting to achieve interoperability could constitute the offence of copyright infringement.

The French, German and Dutch laws have incorporated the provisions of Article 6 of the *Computer Program Directive* in an almost literal manner and thus

national interpretations do not diverge from the European orientations defined here.

Therefore, emulating software is in principle permitted under the relevant national laws and subject to the limitation to research, to achieve interoperability between the emulation platform and the Multimedia Work. However, reproduction of a substantial quantity of lines of code or of the structure of decompiled computer programs is not allowed.

8. LEGAL IMPACT ON EMULATION OF HARDWARE

Within the context of emulating hardware several areas were investigated. Each of these are explained in the following sections.

8.1. Patent Protection

In general, hardware components are often patented. A patent is (a set of) exclusive rights granted by a patent office to an inventor for a limited period of time. In turn, a public disclosure of an invention is given which allows the rights holder to gain a benefit from the invention in competition with other vendors on the market. As far as patent rights are concerned, it is necessary to distinguish the following situations:

- If hardware is not protected by any patent rights, there are no restrictions to undertake reverse engineering necessary to emulate the applicable hardware and use the resulting emulation program;
- If hardware is protected by patent rights which are still in force, it is not allowed to carry out the activities described above, depending on what is specifically claimed in the patent. This needs to be assessed on a case-by-case basis;
- If hardware was protected by patent rights which are no longer in force, there are no restrictions.

French, German and, to some extent, Dutch law limit patent protection as it does not extend to private or non-commercial purposes (private use exception) and does not extend to acts solely intended for research or testing on the patented subject matter (experimentation exception). However, if emulation is meant for giving access to digital objects to the public, none of these exceptions are applicable. So, KEEP cannot make use of emulators that mimic hardware still protected by a patent without asking permission of the rights holders beforehand.

All European countries' domestic laws and regulations protect national patents for twenty years (maximum) from the filing date. The filing of a patent often occurs several years before the marketing of the related invention. In addition, to keep the patent in force, annual renewal fees must be paid to the different national intellectual property's Offices where patents have been filed.

Consequently, in most cases, the patented product or process will become public domain before the term of the twenty years protection. It is however necessary to verify, on a case-by-case basis, for each invention identified, whether the patent rights are still in force.

Once the hardware protected by patent rights falls in the public domain, the related invention is free for exploitation. Therefore, emulation of older computer hardware (older than twenty years) is likely to be permitted.

8.2. Emulation of Semi-Conductors (Computer Chips)

Semi-conductors can be protected by patents in relation to their hardware layer and by copyright when it relates to the firmware (software) layer they may embed. In addition, semi-conductors enjoy a special protection as far as their topography or mapping is concerned. In this case, the rules deriving from Council Directive of 16 December 1986 *on the legal protection of topographies of semiconductor products* (87/54/EEC) [5] should be considered.

The directive provides protection to the 'topography of a semi-conductor product' in so far as it satisfies the conditions "that it is the result of its creator's own intellectual effort and is not commonplace in the semi-conductor industry" (article 2.2). In such a case, the rights holder can forbid the reproduction of the topography by others. This rule, however, carries exceptions that seem relevant to emulation research:

- a Member State may permit the reproduction of a topography privately for non-commercial aims (article 5.2);
- The exclusive rights granted to the rights holder shall not apply to the sole reproduction for the purpose of analyzing, evaluating or teaching the concepts, processes, systems or techniques embodied in the topography or the topography itself (article 5.3);
- The exclusive rights referred to in the first paragraph shall not extend to any such act in relation to a topography created on the basis of an analysis and evaluation of another topography, carried out in accordance with Article 5.3 (article 5.4).

It is quite likely that activities regarding emulation with respect to the reproduction of semi-conductor chip masks, if any, would fall within the scope of those exceptions.

8.3. The Use of Emulated Hardware From Third Parties

Within the emulation community a lot of third party emulation software is already available under either an open source or a proprietary license. After verifying whether such third party emulators are not infringing the rights of the emulated environment right holder, such software may be used within the limits of their licenses

(irrespective of their Open Source or proprietary character) and of the applicable law. With respect to Open Source licenses most of them indeed allow the use and modification of source code, and permit further distribution of the resulting product, even as commercial distribution. As such, emulator developers could use existing Open Source licensed code in their own emulator. Most Open Source software licenses require that the recipient of each resulting product must be given (a) access to the source code and (b) a license to the product which is in line with the initial Open Source license. In other words, the company or heritage institution/KEEP partner that incorporates the Open Source licensed code into its own software cannot distribute it more restrictively than the initial Open Source license. As a result, the licensed code remains 'free', even when embedded into future derivative works.

9. RECOMMENDATIONS & FUTURE WORK

Based on the analysis presented in the previous sections a couple of recommendations can be drawn.

9.1. Regarding Media Transfer

It is reasonable to consider that the legal risk of transferring data from old carriers is relatively limited as long as conservation is only done at cultural heritage organisations and access is only granted on small scale to individual researchers. However, none of the research exceptions are applicable if the emulation platform is meant to be made available to the public at large to give them access to the digital objects, as is KEEP's goal.

Therefore, it is strongly recommended that copyright law is adapted to fit the Information Technology age in which we live in. Today, several Digital Rights Management (DRM) mechanisms have been developed and applied to regulate the usage of digital content. For such content it is common practice that the use restrictions are transferred together with the object itself or that restrictions in usage and/or transfer of the Multimedia Works are encoded within the digital object. Therefore, copyright law needs to be adapted to focus more on protecting against the unauthorized usage of Multimedia Works rather than just simply prohibit to transfer Multimedia Works.Especially for those media types which are already obsolete or becoming so, a general exception for cultural heritage institutions to transfer digital media carrier for archival and access purposes without any technical restrictions is urgently needed on both national and European levels.

Two possible solutions called *the legislative path (9.2.1)* and *the negotiation path (9.2.2)*.

9.2. Regarding Emulation of (Embedded) Software

As explained before, the EU only allows reproduction to carry out decompilation (reverse engineering) for interoperability purposes provided that the resulting program does not incorporate portions of the code that is subject of decompilation. The practical steps that should be taken to mitigate the risks of copyright infringement are twofold:

9.2.1. The Legislative Path

A consortium of stakeholders could launch an initiative aiming at modifying the current *Community Framework* of the EU and more specifically the *Computer Programs Directive* to include a new exception. Such an exception should allow cultural or legacy institutions such as national libraries, archives and museums to perform the necessary steps for reproduction in order to preserve Multimedia Works running on proprietary programs, through emulation or any other relevant technique. As a condition to this exception, it could be envisaged that the institutions would only be allowed to do so when such proprietary programs are no longer on the market or supported by the relevant manufacturers.

This approach is obviously the most effective one in terms of result and legal security. However, it is likely that software companies will launch a strong lobby to hamper the achievement of such an initiative as they did in the past in respect of the *European Computer Programs Directive* [1] and, especially, against the decompilation exception. As a result, even in the event that the stakeholders would be able to find heavy political support to endorse the initiative, the timeframe necessary to reach the final objective will be long and even longer considering the implementation required at national level.

9.2.2. The Negotiation Path

An alternative solution would be to approach the right holders in order to obtain the required authorisation to proceed with the activities of emulating software. This would require the stakeholders to disclose the existence and purpose of the research to such right holders which may eventually perceive it as a potential threat for their proprietary rights. They may fear the dissemination of a groundbreaking and far-reaching emulation technology. Negotiations are therefore likely to be difficult and, in any event, lengthy, as the rights holders will thoroughly assess the risks and proceed very carefully. Moreover, not all the rights holders will be found as some will have gone out of business.

A possible approach would be to involve those manufacturers as sponsors within the group of emulation stakeholders, by associating their names with the research and enabling them to communicate about their participation in a project of public interest. For instance, Microsoft could see a PR benefit from showing that it is collaborating with national libraries for non-profit purposes. However, it may be difficult to attract the most important software manufacturers within an enlarged emulation community.

It is also highly probable that those manufacturers would require an insight into the emulation technology which may raise concerns in terms of protection of

intellectual property rights developed as part of the emulation research.

9.3. Regarding emulation of hardware

The first practical step that should be taken with respect to emulating hardware would be to verify whether the processes or products to be reverse engineered and emulated are protected by patent claims still in force. This should be done on a case-by-case basis with support from patent agents.

If the conclusion is that the processes or products are not, or not any longer, protected then no obstruction exists to undertake the emulation operations and develop the corresponding emulator. However, if the processes or products are still protected by patent claims, then the two options described in the previous section (*legislative* or *negotiation path*) seem the only two options.

In this regard, the legislative path is probably a more difficult option than for software as there is no unified European framework governing patent protection that could be amended. Furthermore, any national or European initiative would probably violate the provisions of the TRIPS agreement [23] which is signed on international level. One can imagine that the United States or East Asian countries would probably strongly disapprove in protection of the interests of their national manufacturers.

Concerning the *negotiation path*, the obstacles mentioned in respect of software are relevant here as well, unless the research community would specifically target their interests in technology that is economically outdated as manufacturers probably do not see any commercial or technological value anymore.

9.4. Future work

Clearly, work has to be done in the field of legislation for preservation and use of digital objects. The current *Community Framework* and national laws do provide some structure but legislation is scattered over many domains and different national interpretations exist. Major legal obstacles still exist regarding copyright protection, software re-use and recreation of computer environments based on patented hardware. National libraries, archives and museums try to fulfil their task in offering long-term access to authentic digital material, but they cannot violate the law.

Therefore, a strong group of stakeholders should be formed representing a variety of cultural heritage organisations such as libraries, archives and museums. Moreover, research organisations and companies could join as each organisation will face severe loss of access to old media and information as long as regulations will not change. A two-fold initiative could be deployed:

- focus on a long-term approach of lobbying for better legislation regarding preservation and use of software, media migration and emulation of hardware;
- focus on negotiations with software manufacturer and hardware manufacturers to find solutions on the mid-term, enabling the emulation of their software and hardware for access to digital information.

Only this way, alignment can be reached between what technically is possible and legally allowed to enable preservation and access of digital information in Europe and beyond. The three national libraries involved in this research are now considering the above-mentioned strategy and invite others to join forces.

10. ACKNOWLEDGEMENTS

The KEEP legal study was subcontracted to the international law firm Bird&Bird. Research has been done in close cooperation with the legal departments of the Bibliothèque nationale de France (BnF), Deutsche Nationalbibliothek (DNB) and the Koninklijke Bibliotheek (KB) – the National library of the Netherlands. Special thanks go to the people involved in this research: Stephane Leriche and Anne-Sophie Lampe (Bird&Bird), Harold Codant (BnF), Anne Baranowski (DNB) and Annemarie Beunen (KB). Also thanks to David Michel (Tessella) for reviewing this paper.

11. REFERENCES

[1] Battle on European Directive Program regarding software, available at: http://www.computerweekly.com/Articles/2005/07/06/210757/european-software-directive-defeated.htm (accessed 11 May 2010)

[2] Dennis W.K. Khong, "Orphan Works, Abandonware and the Missing Market for Copyrighted Goods", *International Journal of Law and Information Technology 2007/1*, available at: http://ijlit.oxfordjournals.org/cgi/content/abstract/15/1/54 (accessed 8 July 2010)

[3] Document presenting the state of legality in the field of computer media copyright, KEEP, January 2010, available at: http://www.keep-project.eu/ezpub2/index.php?/eng/content/download/5329/26645/file/KEEP_WP1_D1.1.pdf (accessed 9 July 2010)

[4] Dutch law on TMP: section 29a subs. 4 of the Dutch Copyright Act.

[5] EC Directive 87/54/EEC, available at: http://europa.eu/legislation_summaries/internal_market/businesses/intellectual_property/l26025_en.htm (accessed 11 May 2010)

[6] EC Directive 91/250/EEC, available at: http://ec.europa.eu/internal_market/copyright/docs/docs/1991-250_en.pdf (accessed 11 May 2010)

[7] EC Directive 96/9/EC, available at: http://ec.europa.eu/internal_market/copyright/docs/

databases/evaluation_report_en.pdf (accessed 11 May 2010)

[8] EC Directive 2001/29/EC, available at: http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2001:167:0010:0019:EN:PDF (accessed 11 May 2010).

[9] EC Directive 2009/24/EC, available at: http://eurlex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2009:111:0016:0022:EN:PDF (accessed 11 May 2010)

[10] Emulation Expert Meeting (EEM) 2006, The Hague, The Netherlands. Available at: http://www.kb.nl/hrd/dd/dd_projecten/projecten_emulatie-eem-en.html (accessed 4 May 2010).

[11] French law on TMP: article 10 of Décret nr 2006-696 of June 13.2006.

[12] Hoeven, van der, J., Lohman, B., Verdegem, R., "Emulation for digital preservation in practice: the results", *Proceedings of the International Conference on Preservation of Digital Objects*, Beijing, China, 2007.

[13] Hoeven, van der, J., Wijngaarden, van, H., "Modular emulation as a long-term preservation strategy for digital objects", *Proceedings of the International Web Archiving Workshop*, Vienna, Austria, 2005.

[14] KEEP project website http://www.keep-project.eu (accessed on 4 May 2010)

[15] OAIS reference model (ISO 14721:2003), available at: http://public.ccsds.org/publications/archive/650x0b1.pdf (accessed 12 May 2010)

[16] PARSE.Insight survey report, December 2009, available at: http://www.parse-insight.eu/publications.php#d3-4 (accessed 10 July 2010)

[17] Planets project, available at: http://www.planets-project.eu (accessed 6 July 2010)

[18] Planets suite, available at: http://planets-suite.sourceforge.net/suite/ (accessed 6 July 2010)

[19] Requirements and design documents for services and architecture of emulation framework, available at: http://www.keep-project.eu/ezpub2/index.php?/eng/content/download/7918/39623/file/KEEP_WP2_D2.2_complete.pdf (accessed 6 July 2010)

[20] R.J. Adair, R.U. Bayles, L.W. Comeau, and R.J. Creasy, "A virtual machine system for the 360/40", *IBM Cambridge Scientific Center report 320-2007*, Cambridge, MA, May, 1966.

[21] Senftleben, M.R.F. (2009). Pacman forever - preservering van computergames. AMI, 2009(6), 221-228.

[22] Specification document for all layers of general-purpose virtual processors, available at: http://www.keep-project.eu/ezpub2/index.php?/eng/content/download/7917/39619/file/KEEP_WP4_D4.1.pdf (accessed 6 July 2010)

[23] WTO TRIPS agreement, available at: http://www.wto.org/english/tratop_e/trips_e/t_agm0_e.htm (accessed 11 May 2010)

THE DEVELOPMENT OF A RETENTION AND DISPOSITION SCHEDULE IN A PRIVATE ENTERPRISE

Ellen Margrethe Pihl Konstad
Det Norske Veritas
Veritas veien 1
N-1322 Høvik
Norway

ABSTRACT

Being a private international enterprise, the ongoing transition from paper to digitally stored documents and records has created some new challenges.

By implementing an Electronic Records Management (ERM) system, the tool for conducting records management is in place, but in order to utilise the possibilities, a revision of the retention and disposition schedule was necessary.

The task of developing a new schedule is time-consuming, but it will be an important tool for future RM work. It gives a good overview of the content of the archives. When implemented it will reduce growth, improve sharing of information and ensure compliance over time. It is also a vital tool for long term planning, in knowing what to keep and for how long, strategies can be developed based on timeframe, cost, need for access and volume. It can also be used in discussions towards historical institutions for the transferee of some or all historic records.

This paper describes the tasks involved in the process towards a new schedule.

1. INTRODUCTION

Det Norske Veritas (DNV) is an independent foundation with the purpose of safeguarding life, property, and the environment. Its history goes back to 1864, when the foundation was established in Norway to inspect and evaluate the technical condition of Norwegian merchant vessels. With 5574 vessels and 230 mobile offshore units in class, DNV is the world's fourth largest class society based on tonnage. In addition to classification, DNV also do certification and consulting services. DNV is located with 399 offices in 100 countries.

As a company, DNV have a 150 year long tradition of keeping information on paper, and along with that experience with evaluating what to keep and what to discard when the information is no longer of any value to the company. Parts of this task have been distributed to the end-user; the true expert knowing the content and the business value of it. In the transition to digital storage and preservation, new challenges have been raised. Since the volume is not physically visible – the end users have not see the same need for disposal of outdated, superfluous and redundant information.

New tools have opened for new ways of working with information in the creating phase as well as new ways of sharing and retrieving information. New techniques have also resulted in new problems in relation to long term preservation. A revision of the retention and disposition plan was needed, based on requirements, routines and the possibilities in a new ERM system. The old plan did not open for different disposition for material on the same entity, e.g. all records related to a project had the same disposition time. The ERM system opened for disposition on document level, enabling a more granular schedule.

2. NEW CHALLENGES

For most countries, the creation and preservation of archives are divided tasks, with a national archive responsible for the preservation and different governmental bodies answering for the creation, where the national archive often is responsible for guiding the creator.[1]

In the transition from paper to digitally stored information, the two tasks of creating and preserving have been merged into one for private enterprises, as part of a document/record life cycle. The ability to access digitally stored information in 40 years or 400 years, meets the same challenges, thereby needing the same strategy for long term storage in addition to plans for what to store.

DNV end user's focus on managing information has also changed. Since digitally stored volumes are not visible in the same way as paper, disposal of this information has not been executed, contributing to a growth rate of 100% every 18 months. In addition, users growing up with the internet and Google, have an expectation of fast and easy access to information.

[1] http://www.lovdata.no/all/tl-19921204-126-002.html#6 Lov om arkiv, § 7. Rettleiings- og tilsynsansvar.

An even faster growth rate, poor quality management of the content and a new search possibility demanded a full review of the way DNV handles its documents and records.

What needs to be preserved, why does it need to be preserved, and what can be discarded? In order to answer these questions, a thorough revision of the retention and disposition plan, including supporting tasks, was initiated.

3. RECORDS MANAGEMENT (RM) GOVERNING LEGISLATION

Private enterprises, at least in Norway, are not governed by local legislation in the same way as public sector.

In the process of revising the retention and disposition plan, few or no national laws governing the creation, retention and disposition of records have been identified, except from financial and human resources related documentation. The general legislation that governs governmental records in Norway focuses on documenting the decisions that have been made and was not transferable to private use.

Not being able to reuse the national legislations and routines, a method for creating a schedule covering the new identified needs was essential.

Standards as ISO 15489 and MoReq2 where investigated.

ISO 15489 states that "Records systems should be capable of facilitation and implementing decisions on the retention or disposition of records"[2] but gives little or no guidance to the content of such a schedule.

MoReq2 identifies the requirements to the REM system[3], but again, no help on forming the schedule itself.

In addition, the National Archive of Norway and the vendor Open Text where contacted, but could offer little assistance.

3.1. Models

In the preliminary work with the schedules, identified internal stakeholders where interviewed. These interviewees were managers and senior/expert users, and had no background in information or records management. Quite a lot of time was spent on RM theory, and a lot of misapprehension arose. In order to avoid this and to visualise RM, the concepts where transferred into simplified models.

3.2. Records Management

The first model is based on ISO 15489, and explains very simplified the difference between a document and a record.

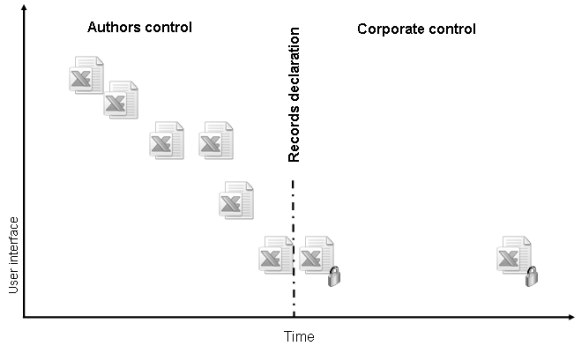


Figure 1. Document vs. records management; the characteristic of a document is that it is under author control where changes to the content, structure and metadata can be made freely within the boundaries of the document management system. When a document has been declared as a record, the control is transferred to the corporation, and the content, context, structure and RM metadata are "frozen".

This model has become the DNV model for Records management and is used in discussions in order to ensure that all participants have the same starting point.

In addition to this main model, 4 sub models where introduced in order to visualise the 4 different lifecycle alternatives that the main model may represent. They also illustrate the difference between declaration time and disposition time trigger.

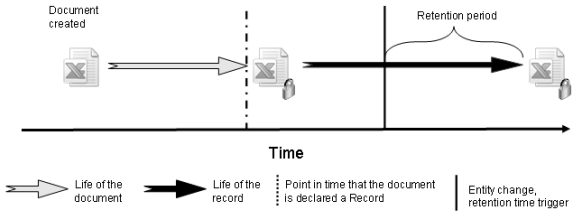


Figure 2. The documents are produced through the production system, and the end user has to manually declare it as a record. This is possible where the end user knows or controls when the document reaches the stage of readiness for declaration, e.g. a final report version.

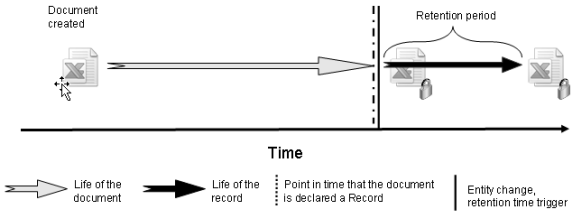


Figure 3. The document is declared a record based on an entity change in the system, e.g. the changing of a status

[2] ISO 15489, part 1:General, 8.3.7" Retention and disposition"

[3] MoRec2 Specifications, Version 1.04, Chapter 5 "Retention and disposition"

from "project active" to "project closed". This is relevant for e.g. project check lists, a document that is being updated during the project, but needs to be declared a record when the project is finished.

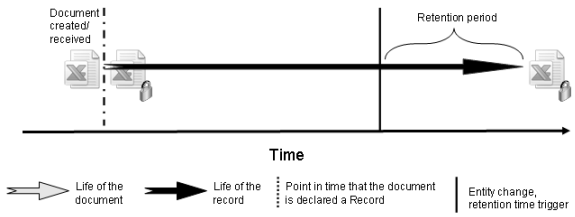


Figure 4. Shows a document that will become a record immediately when received by DNV, e.g. an e-mail from a customer. Record declaration is made when the document is imported into the production/ERM system. It is vital that it is not possible to tamper with the e-mail in the transmission process from the mail system to the production/ERM system to ensure the records authenticity.

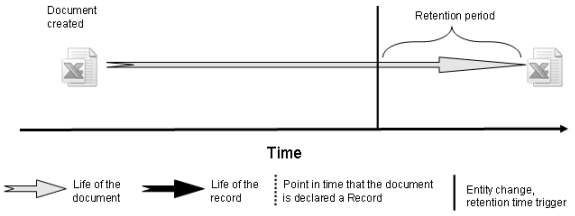


Figure 5. The document is never to be declared a record, and stays in the system as a document.

In combination with these 4 lifecycle scenarios there are different disposition possibilities. The different possibilities that will be implemented in DNV are:

- Automatic disposition of both document/record and metadata stubs
- Automatic disposition of document/record keeping metadata stubs
- Documents/records up for deletion are sent for review

The revision might be performed on document level or for entire entities, e.g. all documents/records belonging to one project.

In addition, some records and documents will for historic purposes be kept permanently.

4. THE PROCESS

First task was to identify "why do private enterprises keep records?"

For private companies, funding of archives has to be justified. Keeping records is an expense. Even for a foundation like DNV, justification has to be identified and accepted in order to receive funding of the archives.

In DNV three reasons for preservation of records have been established:

Figure 6. DNVs model for keeping records.

The core represents records that need to be kept in order to fulfil legal requirements for businesses. This is mostly records related to HR and accounting/finance. (The challenge here is how to be compliant in 100 countries. The retention time varies from 0 to 70 years, with some that we are prohibited to keep for longer.)

The next level is records that are kept for business reasons, e.g. information considered vital for re-use or proof of conduct, because the records information content is allowing the business to run more effectively and efficiently or simply because our customers expect it, in some cases through formal agreements.

The last reason to keep records is for historic purposes. These are records kept in order to document historic events, products or processes. In DNV these records are predefined and approved by our CEO. Documents belonging to this category are typically recurring records as annual reports, development plans and minutes of meetings from board meetings. This category also includes records from major incidents like the Alexander Kielland accident in the North Sea[4], or the records concerning the royal yacht "Norge". Incident records are approved continuously by the owner of DNVs historic archive.

Documents that do not fit into any of these 3 categories, are considered unsuitable as records, and should therefore remain as documents and be disposed of according to the disposition rules for documents.

As part of the work on records, a retention and disposition schedule for documents were also developed in order to automatically discard superfluous information and to avoid a situation where documents are 'kept forever' while records were managed and disposed of.

4.1. Document Types

After identifying and establishing the rules for which records to preserve, the mapping of the different types of

[4]Alexander Kielland was an oil production platform that sank in 1980.

documents existing in DNV were initiated. At present 44 different document types (doc.type) are identified.

All record types have a corresponding document type, but not all document types become records types.

In this process, 27 different synonyms to the type "Agreement" were discovered only in English. The task of translating this into local language has not started, but through implementing doc.types users have the possibility to search for "Agreement" and get hits in local language.

4.2. Process Analyse

In order to really understand which documents where produced in DNV, and which needed to be declared as records, a thorough analysis of our production systems were initiated. This is still work in progress, but 3 of 4 major systems have been completed, analysing each step of the processes, what is input and output. At present, SharePoint 2010 is under implementation, and part of the implementation project is to do a similar analysis.

4.3. Retention and Disposition Schedule

After having established the criteria for which documents that are to become a record, the definition of document types and the process analysis, the concrete work on the retention and disposition schedule could commence.

In order to take full advantage of the ERM system and identify roles that could be governed over time, the schedule ended up with 14 different entities for each rule.

1. Process: DNV core business processes and support processes where records are produced or received.
2. Record owner/responsible: All records and record series shall have an identified owner. The owner can delegate the job to an identified role in DNV.
3. Record identifier: A record may be identified by its correlation to other records or by its content.
4. Record series: A group of identical or related records that are normally used and filed as a unit, and that permit evaluation as a unit for retention scheduling purposes.
5. Document type: The content of the record - what the record is about.
6. Retention purpose: Records retention classified as;
   i. LE- Legal
   ii. BU – Business
   iii. HI – Historic
7. Warrant: Exact reference, including version/edition, to regulatory document (law, rule, regulation, governing document) in which the retention or disposition requirement is stated.
8. Retention period: The period of time the record must be kept before it can be destroyed.
   If the record is to be kept forever, this is to be indicated by using the term "Permanent" instead of stating the number of years.
9. Retention trigger: The trigger for when the retention period starts running.
10. Disposition rules: Rules of disposition action.
11. Storage media: The medium in which the record is kept and managed.
12. Storage facility: The name of the application and / or the physical archive in which the records are stored during the retention period, e.g. Livelink, DNV Historic archive.
13. Outsourcing of the storage facility may occur, but only after an analysis of the rules governing the records. For HR related records, legal counsel must be obtained prior to outsourcing if records are to be stored in another country than the country where the record originated.
14. Security classification: Identification of the level of protection required for the content type.

5. IMPLEMENTATION

Up till 2008 DNV had 4 major production systems with document management functionality, but with no or poor records management functionality. These systems acted as digital information silos, with no exchange of information between the systems. A growing focus on sharing and reuse of information resulted in a major merge project, where files from the different systems where moved into one common repository; Open Text's LiveLink (LL).

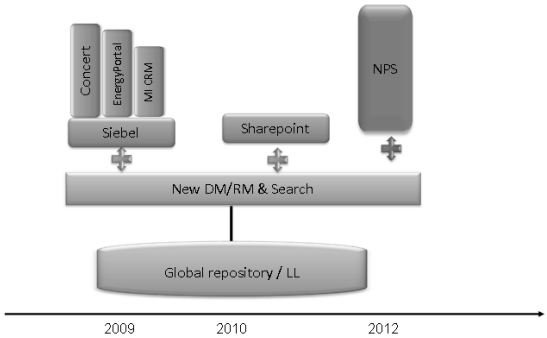


Figure 7. The Conceptual design of the merge project.

With its records management functionality, it has enabled DNV to implement the retention and disposition rules. A "declare records" functionality has also been implemented in the production systems. This combination ensures that DNV's records and documents are managed in a satisfactory manner.

There is no local records management role in DNV, requiring the system to do as much as possible back

office in order not to impose too many new tasks on the end user. One of the back office functionalities implemented is a link between templates and doc.type. In addition all the retention and disposition rules are applied to each document and record on creation.

So far 3 of 4 systems with a common document repository of a total of 4.000.000 files/1255 GB have been merged. Plans are to move the last system in 2012, currently consisting of 8.417.984 files/2750 GB.

6. LESSONS LEARNED

For the end user, the merge of the file repository has together with the implementation of a common search functionality resulted in easier access to the information in DNV. In addition, corporate naming conventions e.g. doc.type have increased the quality of retrieval and enabled search across languages. The manual declaration function is a functionality the end user has been requesting.

A clearer definition of ownership and systematic work towards external legislation has resulted in better governance and compliance with internal and external rules and legislations.

Through the work and the use of simplified models, the general records management maturity in DNV has risen.

The retention and disposition schedule gives an easy overview of how long a document or record needs to be kept and allows for more systematic work towards the objects that needs to be kept for more that 10-15 years. LiveLink supports both migration to a preservation format and differentiated storage media, and strategy work on this topic is currently ongoing.

Generally, the quality of our repository will become better through the declaration functionality and the automatic disposition of documents and records.

7. CONCLUSION

In the transition from paper to digitally stored information, new rules for retention and disposition must be developed in order to utilise the possibilities in the ERM system.

The regulatory landscape international enterprises exist in, arises challenges for the handling of documents and records. Changes in national laws as well as contradictory rules and regulations between countries that the company is represented in and no common “world overview” of which laws that applies, highlight the importance of thorough work towards a common retention and disposition schedule.

It is crucial that enterprises have enough resources and insight to make the right decisions at the time of record declaration and thereby ensuring the correct management and trustworthiness of records through their lifecycle.

Not all private enterprises have the funding to preserve records for historic purposes. In order to preserve the memory of private enterprises, national archive institutions must show initiative in order to preserve this part of history.

There is a need for usage of internationally accepted standards and auditing regimes in private enterprises in order to address and act within the difficult international regulatory area of preserving records.

Session 4b: Preservation Services

SHORT PAPER

TRANSFER AND INVENTORY SERVICES IN SUPPORT OF PRESERVATION AT THE LIBRARY OF CONGRESS

Leslie Johnston
Library of Congress
National Digital Information Infrastructure
and Preservation Program

ABSTRACT

The digital content lifecycle is generally understood as set of activities: select, get and/or produce, prepare and/or assemble, describe, manage, and, as appropriate, make available. At the bit level, digital content is viewed as files on a file system. Many crucial activities of the digital content lifecycle are therefore undertaken primarily at the bit level, including transferring, moving, and inventorying files, and verifying that files have not changed over time. The identifiable entities at the bit-level - files and directories - are widely and easily understood by Library of Congress digital collection data managers and curators. As part of its initial development in support pf preservation services, the Library is working on a suite of solutions to enable the activities of the digital lifecycle for files and directories. Current and planned tool and service development focus on the BagIt specification for the packaging of content; the LC Inventory System to record lifecycle events; and workflow tools that leverage both. The outcomes for the Library include the documentation of best practices, open source software releases, and support for a file-level preservation audit.

1. INTRODUCTION

For the past three years, the Library of Congress National Digital Information Infrastructure and Preservation Program (NDIIPP) and Repository Development Center have been implementing solutions for a category of activities that we refer to as "Transfer" [3, 6]. At a high level, we define transfer as including the following human- and machine-performed tasks:

- Adding digital content to the collections, whether from an external partner or created at LC;
- Moving digital content between storage systems (external and internal);
- Review of digital files for fixity, quality and/or authoritativeness; and
- Inventorying and recording transfer life cycle events for digital files.

The work on transfer has focused primarily on work with external partners, including those that are part of the National Digital Information Infrastructure and Preservation Program NDIIPP[1] [1]; the National Digital Newspaper Program (NDNP)[2] [4, 5]; the World Digital Library (WDL)[3]; and the Library's Web Archiving initiatives.[4]

The development of transfer services is not surprisingly closely linked with bit preservation, as the tasks performed during the transfer of files must follow a documented workflow and be recorded in order to mitigate preservation risks. The goal of bit preservation is to ensure that files and their vital contextual file system hierarchies are retained intact throughout the digital life cycle.

The digital content lifecycle is generally understood as a set of activities: select, get and/or produce, prepare and/or assemble, describe, manage, and, as appropriate, make available. At the bit level, digital content is viewed as files on a file system. Many crucial activities of the digital content lifecycle are therefore undertaken primarily at the file system and bit level:

- Transferring digital files to the control of the appropriate division or project at the Library, whether from external partners or produced internally;
- Moving digital files between storage systems, including archival storage systems;
- Inventorying digital files; and
- Verifying that the digital files have not changed over time.

The identifiable entities at the bit-level – files and directories – are widely and easily understood by Library digital collection data managers and curators. What we call the Content Transfer Services provide a suite of tools and services to enable the activities of the digital lifecycle for files and directories. Many of the existing tools and services have been or are being

[1] For information on NDIIPP, please see: http://www.digitalpreservation.gov/
[2] For information on NDNP, see: http://www.loc.gov/ndnp/ and http://chroniclingamerica.loc.gov/.
[3] For information on WDL, see: http://www.wdl.org/.
[4] For information on the Library's web archiving activities, see: http://www.loc.gov/webarchiving/.

extended to provide additional support for bit preservation activities.

2. CURRENT SERVICE COMPONENTS

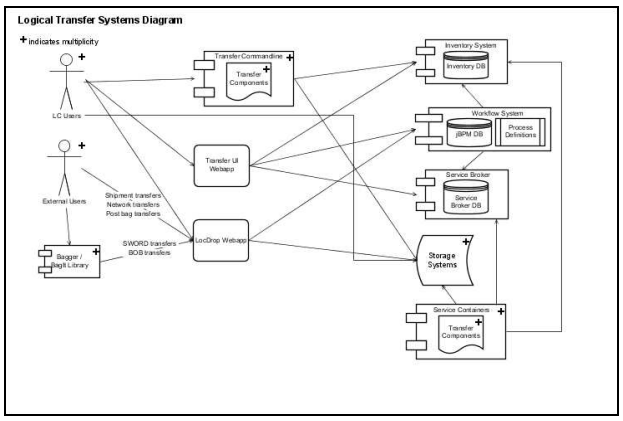


Figure1. Current Library of Congress production Content Transfer Services

BagIt is a specification for the packaging of content for movement between and within institutions. Its package-level metadata and manifest of files and fixities can aid in preservation over time.[5] The base directory of a Bag contains a bag declaration (bagit.txt), a bag manifest (manifest-*algorithm*.txt), a data directory (/data), and an optional bag information file (bag-info.txt). The bagit.txt file is a required file, and simply declares that this is a Bag, and which version of the specification it complies with. The bag-info.txt includes information on the Bag, including descriptive and administrative metadata about the package (not the package contents), as well as the bagging date and human and machine-readable Bag size.

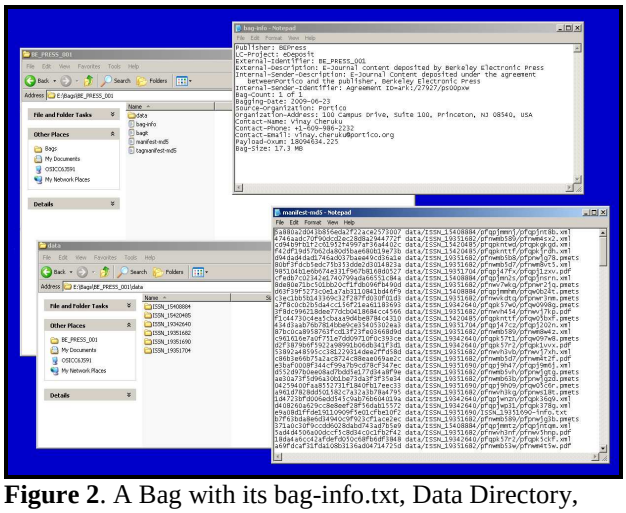


Figure 2. A Bag with its bag-info.txt, Data Directory, and its Manifest.

The manifest lists the names and checksums of the content files; there is an additional checksum manifest for the shipping files. Any commonly recognized checksum algorithm can be used to generate the manifests, and must be identified in the name of the manifest file. The files comprising a package may be transferred in a container format such as ZIP or tar to be unpacked upon receipt. There is also the concept of a "holey" bag, which has the standard bag structure but its data directory is empty. The holey bag contains a "fetch.txt" file that lists the URLs of the content files to be fetched (so-called "holes" to be filled in). Transfer processes follow the URLs, download the files and fill the data directory. The sender's source files do not need to reside in the same directory or on the same server. The content manifest does not obviate the need for descriptive metadata being supplied by the package producer. The manifest assists in the transfer and archiving of the package as a unit, rather than supplying any description of the content.

The data directory is required, and contains the contents of the package, as defined by its producer. The data directory must always be named "/data," and may have any internal structure; there is no limit on the number of files or directories it may contain, but its size should make practical transfers easier, based on physical media limitations or expected network transfer rates. There is no limit on the number of files or directories this directory may contain, but its size should make practical transfers easier, based on physical media limitations or expected network transfer rates. In the Library's experience, 500 GB is the recommended maximum size, although Bags as large as 1.8 Tb have been transferred.

BIL is a Java library developed to support Bag services. A barrier to uptake of the BagIt specification was the inability to automate the Bagging process or support the development of tools. BIL is scriptable and can be invoked at the command line or embedded in an application. It supports key functionality such as creating, manipulating, validating, and verifying Bags, and reading from and writing to a number of formats, including zip, tar, and gzip tar. BIL also supports the uploading of Bags using the SWORD deposit protocol[6] using the Library's extension, BOB (Bag of Bits).

While BIL proved vital in the development of scripted processes, the majority of its potential users at the Library are data managers and curators who are not accustomed to working at the commandline or writing programs. A graphical desktop application for the bagging of content is nearing completion of its development and testing. **Bagger** is a Java application developed on top of BIL with Spring Rich Client[7] as the MVC framework, and a HSQLDB [8] in-memory database. It is implemented as both a Java Webstart

[5] The BagIt specification is available at: http://www.digitalpreservation.gov/library/resources/tools/docs/bagitspec.pdf

[6] For more information on SWORD, see http://www.swordapp.org/.

[7] http://www.springsource.org/spring-rcp

[8] http://hsqldb.org/

application for use across platforms and as a standalone version with its own bundled, Java JRE and various checksum generators. Bagger is a small application, taking up less than 100 MB, and is fully self-contained and requires no administrative privileges or an installer. The limits of its use are the available disk space and memory of the machine where it is used.

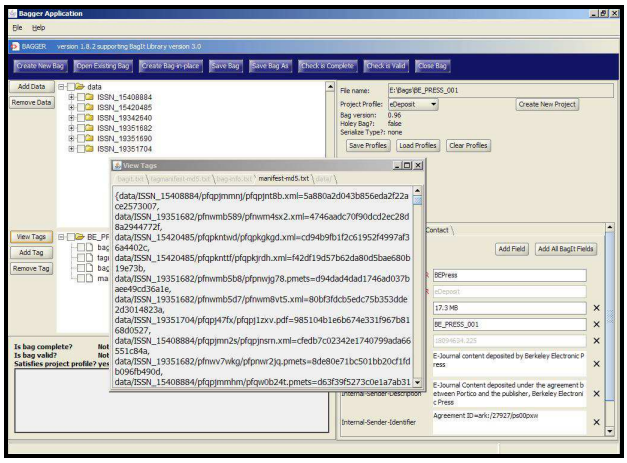


Figure 3. The Bagger Tool for Creating a Bag and its Fixities

The Library has developed utility scripts that support the BagIt specification. The **Parallel Retriever** implements a Python-based wrapper around wget and rsync, and transfers Bags and fills Bags when given a holey Bag manifest and a fetch.txt file. It supports rsync, HTTP, and FTP protocols. The **Bag Validator** Python script checks that a Bag meets the specification: that all files listed in manifest are in the data directory and that there are no duplicate entries or files that are not listed in the manifest. **The VerifyIt Shell** script is used to verify the checksums of Bag files against its manifest. These scripts and the BIL Java Library have been released as open source on SourceForge.[9] Bagger is the next tool under review for open source release.

The **Inventory System** keeps track of and enables the querying of important events in the preservation lifecycle of a Bag and its contents. Its data model is implemented using Java objects mapped to a mySQL database using Hibernate [10] for object-relational mapping. The goal in developing the Inventory Service is to satisfy needs identified through the process of doing transfers and attempting to record their outcomes as well as track the files once their enter the Library's infrastructure. These needs include keeping track of package transfers for a project, tracking individual packages and life cycle events associated with them, and a list of the files that make up each package and their locations. For legacy collections these tools can be pointed at existing directories to package, checksum, and record inventory events to bring the files under initial control. The data in the Inventory System can be used as a source to generate PREMIS metadata[11].

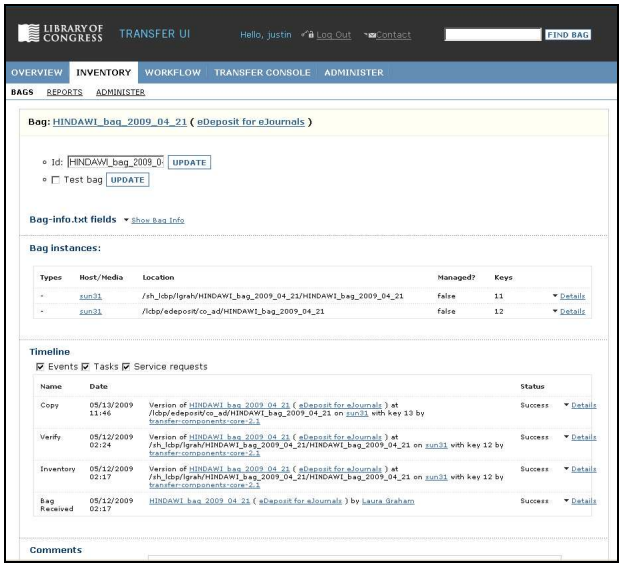


Figure 4. Reviewing the Life Cycle History of a Bag in the Inventory System

Packages are associated with a program and/or project, which are associated with a custodial unit, a content type (textual, still image, audio, etc.), a content process (partner transfer, digital conversion, web archiving, etc.), and an access category. Since it must also represent the history of a package, it records location paths and events that occur on a package level and on a file level. Examples of events include:

- Received Events, which include initial checksum verification and recording into the inventory;
- Quality Review Events, recorded when quality review is performed and noted as passed or failed;
- Accepted or Rejected Events, recorded when a project accepts or rejects curatorial responsibility for a package, usually due to verification failure or a failure to meet expected standards;
- Copy or Move Events, recorded when content is copied or moved from one location to another;
- Modification Events, recorded when a package or file has been modified, added or deleted;
- Delete Events, when entire packages are removed from the system;
- Ingest Events, when content has been ingested into a repository or access application;
- Recon Events, for the inventorying of legacy content already under Library control; and
- Verify Events, for ongoing auditing of fixities.

All events are recorded with the name of the performing agent and full date/timestamps. Multiple copies of content can be recorded as related instances, each with their own event history.

[9] http://sourceforge.net/projects/loc-xferutils/
[10] https://www.hibernate.org/

[11] http://www.loc.gov/standards/premis/

The Library has implemented low-level services such as file copying, inventorying, and verification, which are distributed across multiple servers as service containers. Mechanisms are available for invoking, managing, and monitoring the services through the command line or a web interface. Of particular note is the Copy Selector, which provides transparent access to a number of supported transfer protocols and tools; depending upon the source and copy locations, the most appropriate mechanism will be automatically selected (rsync, SCP, Signiant[12]) without the user having to be aware of the best option. Inventorying and verification services take advantage of the BIL Library and the Inventory Services.

The **Transfer Console**/UI is a web application that provides access to most aspects of the above services, plus project-specific workflows. It allows viewing and updating of the Inventory System, ad hoc transfer services (the Transfer Console), the monitoring and management of transfer services and workflows, as well as auditing and reporting functions. The name of this service is somewhat misleading; while it originally supported only transfer functions, it has been extended to supporting auditing and reporting on all inventoried content in the Library's server environment. The Transfer Console UI was implemented using Spring MVC.[13]

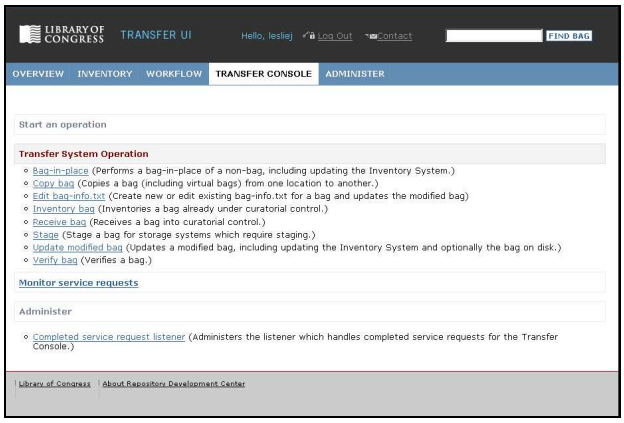


Figure 5. Transfer Console Functions

The **Workflow Framework** supports the implementation of project-specific workflows that automate parts of the digital lifecycle by coordinating machine and manual processes. The underlying workflow engine is jBPM, an open-source workflow system. [14] The drivers of a workflow are process definitions, which represent the process steps. jBPM Process Definition Language (jPDL), the native process definition language of jBPM, is used to encode the workflow process steps as XML. A workflow can be designed using the visual editor Graphical Process Designer, a plug-in for the Eclipse platform.

[12] http://www.signiant.com/
[13] http://www.springsource.org/
[14] http://www.jboss.com/products/jbpm/

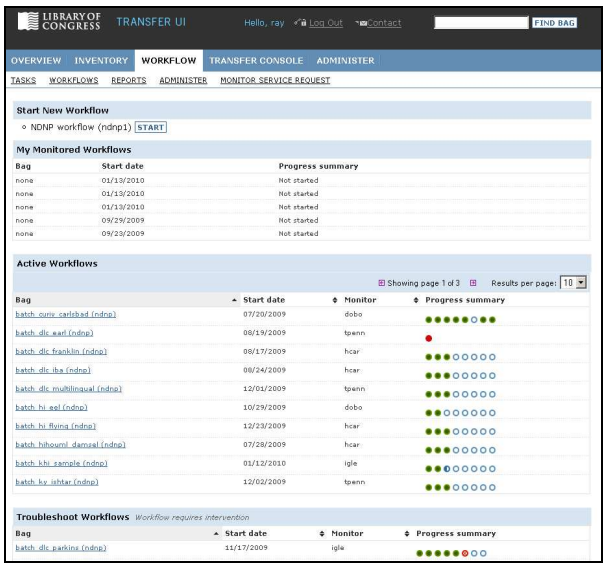


Figure 6. Overview of Workflows for the Processing of Batches

In order to support the expanding numbers and types of transfers, a tool was needed to help automate transfers. The **LocDrop Service** is a web-hosted application for use by transfer partners in registering a new transfer; this application will support the registration and initiation of the transfer content via network transfer and via fixed media, such as hard drives or DVDs. LocDrop uses SWORD as its deposit protocol. At the time of this writing, LocDrop is in its initial use by multiple Library digital content acquisition projects, incorporating feedback into continued development.

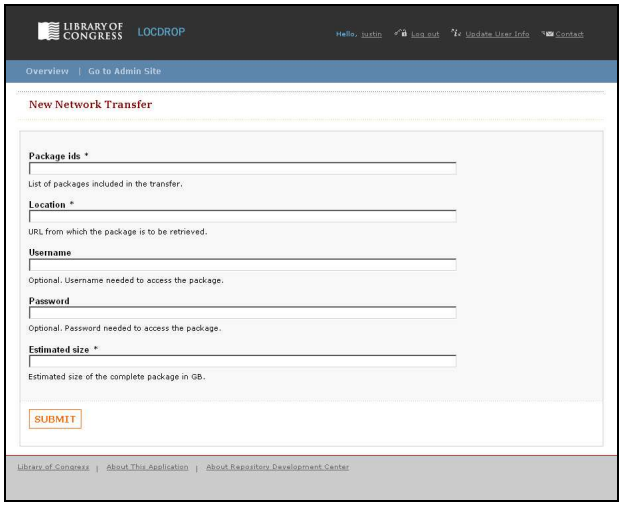


Figure 7. Initiating a New Network Transfer Using LocDrop

All of the applications are developed using an agile process, and undergo extensive QA testing at the completion of each iteration. As each application nears a state where it is feature complete, it is released to staff from one or more projects for user acceptance testing,

the results of which are incorporated into the development process. After testing by partners, a number of new features were identified for inclusion in LocDrop. While the Inventory System and the Transfer Console started out as two applications, user feedback showed that it would more useful to integrate the two services into a single interface. And as a result of testing by internal Library users and partners, the interface for Bagger changed significantly from the 1.x to the current 2.x development versions.

3. PLANNING FOR FUTURE WORK

At the time of this writing, the Content Transfer Services have been put into production for NDNP [6]. The Inventory Tool has been put into production for content transferred to the National Digital Information Infrastructure and Preservation Program (NDIIPP), and production implementation is nearing completion for the Inventory, LocDrop, and Bagger applications.

The development of these services is ongoing at the Library, tentatively scheduled through 2011. A number of tasks have been identified for the remainder of the initial period and transition into full production. An inventory that is independent of any storage system allows the Library to track the location of digital content and checksums to support auditing. While procedures for inventorying newly acquired or produced Bags/digital content is in place for some projects, procedures must be put in place for inventorying all new Library content. As well, a complete inventory of all existing legacy content and full coverage of the production Library server environment is required. This effort is underway.

Currently the Transfer UI and Transfer Console support a workflow for the National Digital Newspaper Program as well as ad hoc and project-specific transfer and inventorying activities. We envision additional project-specific workflows can and will be developed using the Workflow Framework and integrated into the UI to automate reliable, repeatable Bag-level bit preservation activities. As program offices/projects identify their needs, workflows will be formalized and added into the framework.

As the tools and services move into production and use by a greater number of Library projects and staff, the interface will require review and revision for increased usability. An ongoing iterative review and revision of interfaces will be put into place.

These services fit into a larger context of development over the next three years at the Library to implement tools that enable staff across the Library to easily perform digital content management and curation tasks. While we are currently focusing on Bag-level bit preservation, not all content will always be Bagged, and data managers and digital curators think in terms of files, not Bags. The current Bag-level services include limited tracking of files within Bags, but do not currently support file-level auditing and reporting other than lists and counts of file format types. The planned progression of work is to complete the development of services supporting Bags and then move on to file-level services. These services will be implemented as extensions to the Inventory System. Once all Bag-level services are in production (inventorying, auditing, and reporting on Bags), work will commence on adding file-level services, such as file format auditing, file validation, and, potentially, preservation risk reporting. This work requires that policies and procedures on preservation storage, auditing, and preservation formats be in place.

When data managers and digital curators think of files, it is often in terms of their relationships to "objects" that they represent and collections that they are part of. We will continue to focus on bit preservation, but we are considering methods to additionally support an overlay of services that identify which files have relationships to each other (compound objects, master and derivatives, etc.), which file(s) represent which objects, and potentially link to descriptive metadata in other systems. Understanding that a file is a TIFF that represents a page from a specific atlas in the Geography and Maps Division, that another file is a JPEG2000 derivative file representing the same page, and that a third file is a JPEG used as a web thumbnail in addition to managing those bits is important for the preservation and sustainability of the collection as a whole.

4. CONCLUSIONS

Why are such transfer tools and processes so important? After much experimentation, the best transfer practices that have emerged relied upon established, reliable tools; well-defined transfer specifications; and good communication between content owner and content receiver. Each transfer provided insight into the developing content transfer best practices and each exchange brought more expertise. The digital preservation community continues to engage with transfer best practices, helping these practices to evolve. Ultimately, these practices and tools focus not just on transfer optimization, but on ways in which to improve the communication between submitter and receiver. The most important part of transfer is not the connection but the exchange of information. Communicating what is coming, when it will arrive, what form it will take, making the process predictable and flexible is vital.

Why are we looking at close integration between transfer and inventory functions? Inventorying and audit functions have been identified as a vital aspect of data curation. Inventory services can bring several benefits, including collection risk assessment and storage infrastructure audits. Realizing any benefits for effective data management relies on knowledge of data holdings. Knowledge of file-level holdings and

recording of life cycle events related to those files from the moment that they enter the collection and in every future action reduces future risk by storing information that can be used in discovery, assessment, and recovery if and when a failure occurs.

Transfer processes are not surprisingly linked with preservation, as the tasks performed during the transfer of files must follow a documented workflow and be recorded in order to mitigate preservation risks. Defining, implementing, and documenting appropriate transfer processes depends on the requirements of each collection building project, which can vary wildly. While our initial interest in this problem space came from the need to better manage transfers from external partners to the Library, the transfer and transport of files within the organization for the purpose of archiving, transformation, and delivery is an increasingly large part of daily operations. The digitization of an item can create one or hundreds of files, each of which might have many derivative versions, and which might reside in multiple locations simultaneously to serve different purposes. Developing tools to manage such transfer tasks reduce the number of tasks performed and tracked by humans, and automatically provides for the validation and verification of files with each transfer event.

Bit preservation is not synonymous with digital preservation, but is rather an essential subset of digital preservation activities. So why is the focus on bit level operations? Bit preservation is not a solved problem [7]. Bit preservation is a useful starting point because bit-level activities tend to have more in common than activities at other levels. The act of copying a file is the same regardless of whether the file is an image or text or geospatial data. All files should have their formats validated and the checksums regularly verified, whether they represent newspaper pages or a photographs or manuscripts. As well, it is often sufficient to guarantee only the preservation of digital content as bits; in some situations that is all that is possible.

The work at the Library described in this paper has not focused on storage systems (as per Rosenthal); that work is progressing in the Enterprise Systems Engineering group at the Library and elsewhere [8].[15] Inventorying and audit functions have been identified as a vital aspect of data curation and preservation. The Library's developing services provide observability of the state and location(s) of files, enabling querying, auditing and reporting.

This allows the Library to manage its bits as well as additional levels of abstraction: that the bits represent certain types of data (file formats), and that they have relationships (to batches, projects, curatorial divisions). Knowledge of file-level holdings and recording of life cycle events related to those files from the moment that they enter the collection and in every future action reduces future risk by storing information that can be used in discovery, assessment, and recovery if and when a failure occurs. This reduction of risk is vital to the Library's near-term preservation activities.

[15] See the presentations from the "Designing Storage Architectures for Preservation Collections" meeting, held September 22-23, 2009, at the Library of Congress: http://www.digitalpreservation.gov/news/events/other_meetings/storage09/index.html.

5. REFERENCES

[1] Anderson, Martha. "2008. Evolving a Network of Networks: The Experience of Partnerships in the National Digital Information Infrastructure and Preservation Program", *The International Journal of Digital Curation* (July 2008: Volume 3, Issue 1). http://www.ijdc.net/ijdc/article/view/59/60.

[2] Beckley, Elizabeth Thompson. "LOC Expands Tech Focus: Saving Sound and Scene", *FedTech Magazine*, 2008. http://fedtechmagazine.com/article.asp?item_id=490

[3] Johnston, Leslie. "Identifying and Implementing Modular Repository Services: Transfer and Inventory*", Proceedings of DigCCurr 2009: digital curation: practice, promise & prospects*: *April 1-3, 2009. ed.Tibbo, ,Helen R. et. al., Chapel Hill, N.C., 2009.*

[4] Littman, Justin. "A Technical Approach and Distributed Model for Validation of Digital Objects", *D-Lib Magazin* , Vol 12, Nr 5, May 2006: http://www.dlib.org/dlib/may06/littman/05littman.html.

[5] Littman, Justin "Actualized Preservation Threats: Practical Lessons from Chronicling America", *D-Lib Magazin,* Vol. 13, Nr. 7/8, 2007. http://www.dlib.org/dlib/july07/littman/07littman.html.

[6] Littman, Justin "A Set of Transfer-Related Services", *D-Lib Magazine*, Vol. 15, Nr.1/2, 2009. http://dlib.org/dlib/january09/littman/01littman.html.

[7] Rosenthal, David S. H. "Bit Preservation: A Solved Problem?" *Proceedings of iPRES2008*, London, UK, 2008. http://www.bl.uk/ipres2008/presentations_day2/43_Rosenthal.pdf

[8] Schiff, Jennifer L. Library of Congress Readies New Digital Archive. *EnterpriseStorageForum.Com*, 2007. http://www.enterprisestorageforum.com/continuity/article.php/3704461

MOPSEUS – A DIGITAL REPOSITORY SYSTEM WITH SEMANTICALLY ENHANCED PRESERVATION SERVICES

Dimitris Gavrilis
Digital Curation Unit, Institute for the Management of Information Systems, Athena Research Centre, Athens, Greece
d.gavrilis@dcu.gr

Stavros Angelis
Digital Curation Unit, Institute for the Management of Information Systems, Athena Research Centre, Athens, Greece
s.angelis@dcu.gr

Christos Papatheodorou
Department of Archives and Library Sciences, Ionian University, Corfu, Greece and Digital Curation Unit, Institute for the Management of Information Systems, Athena Research Centre, Athens,
papatheodor@ionio.gr

ABSTRACT

Repository platforms offer significant tools aiding institutions to preserve the wealth of their information resources. This paper presents the data model as well as the architectural features of Mopseus, a digital library service, built on top of Fedora-commons middleware, designed to facilitate institutions to develop and preserve their own repositories. The main advantage of Mopseus is that it minimizes the customization and programming effort that Fedora-commons involves. Moreover it provides an added value service which semantically annotates the internal structure of a Digital Object. The paper focuses on the preservation functionalities of Mopseus and presents a mechanism for automated generation of PREMIS metadata for each Digital Object of the repository. This mechanism is activated whenever an object is modified and is based on a mapping of the Mopseus data model to the PREMIS data model that ensures the validity of the transformation of the information stored in a Mopseus repository to semantically equivalent PREMIS metadata.

1. INTRODUCTION

Nowadays there exist several platforms that support the development of digital repositories, but a few of them focus on preservation and facilitate the repository administrators to implement preservation plans. On the other hand the existing preservation platforms, such as CASPAR [8] and Planets [9], provide infrastructures to meet the requirements for preservation actions of large memory organizations such as national libraries and archives. A crucial issue is how much effort users are required to put in order to develop digital repositories on top of such platforms, especially when these users are small institutions with tight, small budgets [15]. Existing repository platforms, such as eSciDoc (http://www.escidoc.org/), offer a number of powerful services, while some, such as Blacklight (http://www.projectblacklight.org/), offer an easy interface and some other, such as RODA [13] (http://roda.di.uminho.pt/), provide preservation features. However they are complex for small – medium organizations and/or demand a number of pre-requisites to be setup.

This paper presents Mopseus, a digital library service, inspired by the conceptualization of [11] and built on top of Fedora-commons middleware that provides repository development and management services in combination with basic preservation workflows and functionalities. These functionalities are based on an infrastructure that semantically correlates the repository content. Mopseus is designed to facilitate institutions to develop and preserve their own repositories [1]. In comparison to the Fedora-commons platform, Mopseus provides a repository system, without the need of customization and the programming workload that Fedora-commons involves. Additionally, Mopseus indexing process is based on a RDMS, ensuring efficiency.

The main objective of the paper is to present an enhancement of the preservation features of Fedora-commons platform implemented by Mopseus. Mopseus is based on an expressive data model aiming to enrich the vocabulary of relationships between the entities of the Fedora-commons model, which are the repository objects and the data structures they contain. The proposed vocabulary revises and improves the existing relationships and defines them explicitly and formally using RDFS. This extension enables the management of information concerning the provenance of the Digital Objects. The new data model is mapped to PREMIS data model [12] in order to automatically generate and incorporate valid PREMIS metadata in Fedora-commons

repositories. Each time a workflow, consisted of a number of events, is carried out and affects the status of a set of repository objects, then PREMIS metadata are automatically generated for each affected object and stored in the repository. Thus the PREMIS metadata generation mechanism is integrated with the Mopseus workflow management component, modifies essentially the logging mechanism and enriches the FOXML [7] schema of Fedora-commons.

In the next section the Mopseus architecture is outlined and its data model and main functional components are presented. In section 3 the main principles on which the preservation features of Mopseus are based as well as the mapping of Mopseus data model to PREMIS data model are presented. Furthermore the implementation of PREMIS metadata generation mechanism is demonstrated. Finally in section 4 the Mopseus innovative features are discussed and in section 5 the main conclusions of the presented effort are sketched.

2. ARCHITECTURE

2.1. Data Model

Mopseus is based on the main Fedora-commons entities which are the digital objects and datastreams and provides an ontology that defines the relationships between them. In particular the content of a Mopseus repository is stored as digital objects, consisting of datastreams, which can be text/xml, text/rdf or binary (see Figure 1). Thus Datastreams can be correlated to form Digital Objects that are structures of data and metadata. Each Digital Object is described at least by a Dublin Core record implemented as a Datastream. Additional descriptive metadata, following any schema, could be incorporated as Datastreams. A new entity enhancing the Fedora-commons conceptual schema is the container. A Container is a Digital Object which aggregates a set of Digital Objects or other Containers. For instance a collection of the PhD theses of a University Department is a Container, which consists of several Digital Objects (PhD theses) and may belong to anotherContainer, e.g. the collection of the University's gray literature.

Each Mopseus Digital Object is an instance of one of the following entities named namespaces:

- config: The configuration of the repository itself is encoded by and stored as Digital Objects of this namespace. This makes Mopseus a self-describing repository, which means that all information regarding the setup of the repository is stored as Digital Object itself and thus is preserved following homogeneous and common preservation mechanisms.Thus, the required knowledge an administrator needs to have in order to configure and maintain the repository is XML.
- cid: This namespace contains Digital Objects that describe Containers. Containers can hold metadata (e.g. DC), binary Datastreams (e.g. a Thumbnail image) and can form any kind of graph through RDF relations.
- iid: This namespace contains all the Digital Objects that carry actual information (items), consisting of Datastreams.
- trm: This namespace contains all Digital Objects that carry terminology information. These Digital Objects are encoded in SKOS and each Digital Object that resides in the trm namespace represents a SKOS concept.

Another significant entity of the model corresponds to the notion of workflow, which is a sequence of states (or events). Each state incorporates a set of basic operations performed on Digital Objects or Datastreams. The descriptions of workflows, states and their basic operations are stored as Datastreams, in the form of XML documents, and they constitute a part of the Digital Objects description in the config namespace.

The Mopseus data model relationships are categorized to the following classes:

- Digital Objects relations: A Digital Object may be correlated to one or more Containers (or other objects) through a set of partitive or membership relations given by Fedora-commons ontology (http://www.fedora-commons.org/definitions/1/0/fedora-relsext-ontology.rdfs) and enriched by the DCTERMS vocabulary for the DC Relation property [6], forming thus a new ontology named RELS-EXT.

Relationship	Domain	Range	Description
isRDF	Object	Data-stream	Denotes that a Datastream of an object is an RDF document.
isThumbnail	Object	Data-stream	Denotes that a Datastream of an object is thumbnail.
isImage	Object	Data-stream	Denotes that a Datastream of an object is an image.
isImageHighDef	Object	Data-stream	Denotes that a Datastream of an object is a high resolution image.
isDocument	Object	Data-stream	Denotes that a Datastream of an object is a document.
isDocumentPDF	Object	Data-stream	Denotes that a Datastream of an object is a document.
isBinary	Object	Data-stream	Denotes that a Datastream of an object is a bitstream.
migratedFrom	Object	Data-stream	When an object is migrated by a repository, a Datastream is generated, holding all information about its provenance.

Table 1. Mopseus Digital Object – Datastream (RELS-INT ontology) relations

- Digital Objects - Workflows: The state of a Digital Object could be modified by a workflow meaning that there exists a correlation between a workflow and one or a set of particular affected Digital Object/objects. These relations are aligned to the vocabulary of PREMIS EventType element [12] and are directly implemented through the Mopseus services.
- Digital Objects - Datastreams: A Digital Object is consisted of one or more Datastreams through a rich vocabulary of relations referred in Table 1. These relations enrich the semantics of Fedora-commons ontology. A crucial relationship for preservation repositories is the migratedFrom which denotes the incorporation of a Digital Object from other repositories.

All relationships are described in RDFS and stored in Datastreams. Specifically, two Datastreams residing in the config namespace have been implemented:

- RELS-EXT. Contains a slightly enhanced version of the Fedora provided RELS-EXTontology for describing relationship types between Digital Objects and Containers, such as isPartOf, etc.
- RELS-INT. Contains an ontology for characterizing the constituent Datastreams of a Digital Object and the relationships between them e.g.isDocument, isThumbnail, etc. Its main classes and relationships are presented in Table 1.

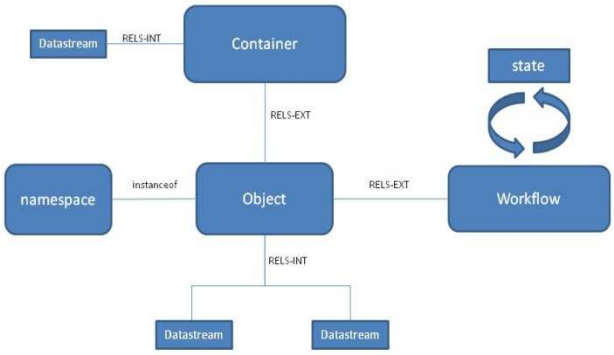


Figure 1. Mopseus data model

2.2. Functional Components

Mopseus consists of two different and distinct parts: a backend which implements the core services of Mopseus (written in Java) and a frontend PHP and JSP API which provides out of the box functionalities that are used to create the different web based GUIs. Only the backend part of Mopseus and certain Fedora-commons functionalities, such as the REST based retrieval mechanisms, communicates directly with Fedora-commons. The main architectural components of Mopseus (see Figure 2) are:

- **Dynamic definition of XML schemas**. Mopseus provides a service for the definition of metadata schemas. The service supports the development of XML schemas, defining the syntax of the metadata elements, their functionality (mandatory/optional elements) and presentation. A new XML schema is automatically transformed into HTML forms and the user can use them to ingest metadata and produce valid XML documents stored as Datastreams. The service that translates the XML schema definition to a working HTML form also supports a number of other features such as: creating an object from templates, creating an object from a mapping mechanism (see below), etc. The metadata schema definitions are stored in the schemas Digital Object of the config name space.
- **Relations manager**. The relations manager allows for the easy management (insert, delete) of relations both external (between objects and Containers) and internal (between Datastreams). The parameters of this service are the Datastream on which a relation should be added, deleted or

modified and the ontology that keeps the Mopseus relationships. The service allows the user to define relations in a flexible manner and use different ontologies on different Datastreams.

- **RDBMS Synchronization**. A mechanism was developed to dynamically synchronize any or all the elements of the hosted XML schemas with an external RDBMS database (currently MySQL is supported). This process features a flexibility which is achieved by automatically mapping XPath Queries to SQL queries. Furthermore, this mechanism can also store in the RDBMS RDF information (e.g. relations) and Datastream information. This process drastically improves the efficiency and flexibility of the indexing of any kind of XML or RDF document stored in Datastreams, makes easier the implementation of a web frontend system and the searching process. All the RDBMS synchronization information are stored in the sync Digital Object of the config namespace.
- **Mapping between XML schemas**. This mechanism allows the mapping between metadata schemas. The mapping is created through an XSLT tansformation. The mapping service can take as parameters a Datastream that contains the XSLT transformation, the Datastream containing the source XML document and the target Datastream. It then can automatically perform the transformation and store the result onto the target Datastream. The XSLT document itself along with the mapping rules are stored in the mappings Digital Object (config:mappings) of the config namespace.
- **Workflow engine**. The workflow engine allows for executing sequences of states, such as ingestion, revision, etc. For each state its input parameters from the previous states and output parameters, which pass to the next states are described. A state invokes a specific service, which in general would be either internal (i.e. one of the mentioned Mopseus components, e.g. a mapping between XML Schemas service) or external (e.g. the use of a data format migration tool, which is not part of Mopseus). Notice that the current version of Mopseus does not support external services.
- **Preservation service**. This service is responsible for the preservation features of Mopseus and provides the following functionalities: (a) maintains a PREMIS log per Digital Object (residing in the PREMIS Datastream) containing all actions and operations that take place on a Digital Object, (b) maintains and checks the checksums for each Datastream and (c) performs simple migrations on binary Datastreams such as PDF documents (this service is currently under development).
- **Terminology service**. The terminology service allows for management of vocabularies, which can then be used in metadata schemas. Information regarding this service is stored in the *terms* Digital Object. (config:terms) of the config namespaceDigital Object. The terms are represented in SKOS.

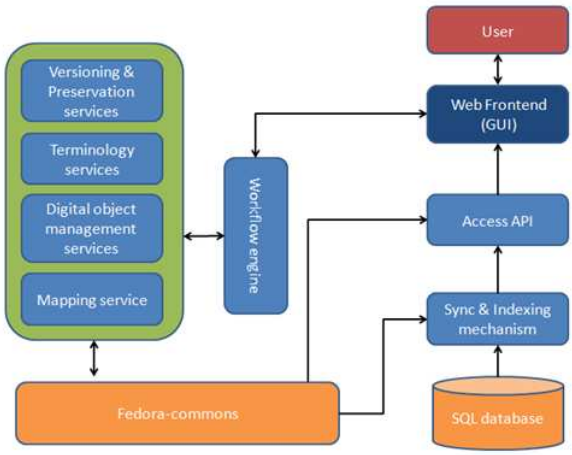


Figure 2. Mopseus architecture. This figure illustrates the basic high-level components of Mopseus and how they inter-operate.

Constructing a user interface in Mopseus is relatively easy. The developer can utilize ready to use components such as an Admin Panel which allows the user to perform operations on Digital Objects, relations and Datastreams. Furthermore, item short views (see Figure 3b) as well as detailed views (see Figure 3c) can be obtained directly from corresponding ready to use XSLT files. Binary Datastreams available for viewing and downloading (e.g. isDocumentPDF, see Figure 3d) can be displayed based on the RELS-INT relations. Finally the Containers an object belongs to can be obtained by the relevant RELS-EXT relations (see Figure 3a).

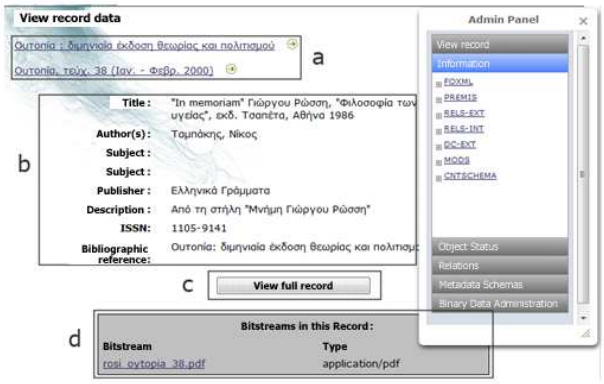


Figure 3. A screenshot of the Mopseus installation at Panteion University.

3. PRESERVATION STRATEGY

3.1. Outline of Strategy

Mopseus preservation strategy follows a set of rules that aim towards a long-term storage and access to the Digital Objects, with respect to small and middle-sized institutions, with a probable low budget. Mopseus is inspired by the OAIS [4] model principles in the sense that (a) the Digital Objects carry meaningful information about their binary content and relationships and (b) this representation information constitutes itself a Digital Object. Thus each Digital Object contains a set of Datastreams and relations. The Datastreams carry both representations of an object and the object's descriptive metadata.

Moreover, Mopseus supports ingestion, access, storage, data management, administration and preservation planning OAIS functionalities. In compliance with the OAIS the Submission Information Packages (SIPs) are transformed to Archival Information Packages (AIPs) with the use of a set of internal Fedora-commons mechanisms. The ingested Digital Objects are checked for integrity, descriptive metadata are generated semi-automatically via the mapping mechanism and preservation metadata are generated automatically in PREMIS. All the Digital Objects are preserved by the Fedora-commons mechanisms, which keep versions of the repository state and content. The versions of a repository are stored internally. Regarding preservation planning, Mopseus provides a migration process from other existing repositories, facilitated through the use of a desktop tool implemented in Java. Currently it supports migration from DSpace repositories. The final Dissemination Information Packages (DIPs) are promoted to consumers through a web-front that selects specific aspects of the Digital Objects to show to the end user.

One of the most representative installations of Mopseus is Pandemos, the digital library of Panteion University, Athens, Greece (http://library.panteion.gr/pandemos). Originally, Pandemos was a DSpace repository, holding approximately 2200 Digital Objects, migrated to Mopseus without any loss of information and at least 5000-5500 new Digital Objects were ingested. For the migration process, the migration tool mapped the DSpace communities, collections and subcollections to the Mopseus Containers by creating the appropriate RDF relationships with the Containers. The original metadata from DSpace were preserved into Mopseus while a PREMIS event was created to indicate the preservation action.

3.2. Mapping the data models

The generation of valid PREMIS metadata presupposes the mapping of Mopseus and PREMIS data models. Mopseus data model was presented in the previous section and PREMIS data model is briefly presented as follows [12]:

The PREMIS data model consists of five entities, according to Figure 4: the *Intellectual Entity* ("a coherent set of content that is reasonably described as a unit"), *Object* ("or Digital Object, a discrete unit of information in digital form"), *Event* ("an action that involves at least one object or agent known to the preservation repository"), *Agent* ("a person, organization, or software program associated with preservation events in the life of an object") and *Rights*, ("or Rights Statements, assertions of one or more rights or permissions pertaining to an object and/or agent"). The Objects are categorized to three types: *file* ("a named and ordered sequence of bytes that is known by an operating system"), *bitstream* ("contiguous or non-contiguous data within a file that has meaningful common properties for preservation purposes"), and representation ("the set of files, including structural metadata, needed for a complete and reasonable rendition of an Intellectual Entity").

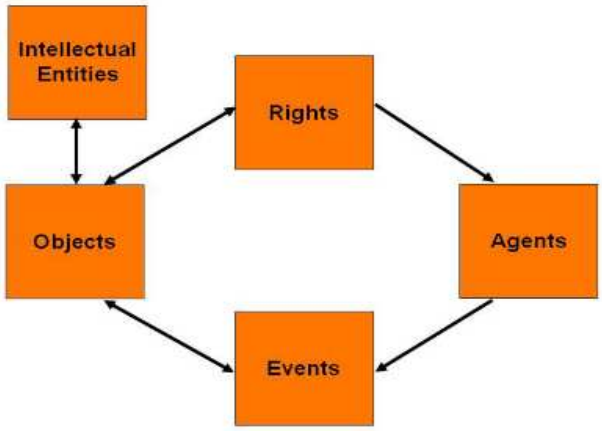


Figure 4. PREMIS data model.

The relationships associate the instances of entities. PREMIS relationships associate the instances of the Object entity as well as the instances of entities of different types. The properties between objects are categorized to three types: *structural*, which are relations between the parts of objects, e.g. the relationships between the files that constitute a representation of an Intellectual Entity, *derivation* relationships, which result from the replication or transformation of an Object, when "file A of format X is migrated to create file B of format Y, a derivation relationship exists between A and B" and *dependency* relationships which "exists when one object requires another to support its function, delivery, or coherence of content". The relationships between different entities are expressed by including in the information for the first entity, a pointer to the second entity.

The mapping of two data models is defined as a sufficient specification to correlate each instance of the

source model with the instances of the target model with the same meaning. The mapping of the two models is presented in Table 2 and analyzed as follows:

The central entity in both models is the Digital Object, though there exist semantic variations between them. A Mopseus Datastream that carries either a binary file or metadata is mapped to the File Category of a PREMIS Object. Moreover a Mopseus Digital Object is mapped to the Representation Category of an Object, since it represents a digital artifact with its binary representation(s) and metadata. Finally the Mopseus entity Container, actually represents a logical aggregation of objects and therefore is mapped to the PREMIS Intellectual Entity, noting that the Intellectual Entity refers to a collection of Objects of the Representation Category.

Mopseus entities	PREMIS entities
Datastream / binary	Object/ file
Datastream / metadata	Object / file
Digital Object	Object /Representation
Container	Intellectual Entity
Workflow / State	Event
Datastream / metadata / Rights	Rights
Datastream (metadata) / Person	Agent
Mopseus relationships	PREMIS relationships
Digital Object - Datastream	Structural relationships
Digital Object - DC Datastream (Fedora-commons default metadata)	Dependency relationships
Digital Object - Digital object (through a Workflow / State	Derivative relationships
Container - Digital Object	Relationships between different types
Workflow / State - Digital Object	Relationships between different types
Workflow / State - Agent	Relationships between different types

Table 2. PREMIS - Mopseus mapping

The Mopseus entity Workflow refers to a sequence of events and thus its subclass State is mapped to the PREMIS Event entity, which represents a particular action in the time-line. The classes Agents and Rights are not expressed explicitly in the Mopseus data model; nevertheless all information that correspond to these entities is stored and available in the Datastreams that hold the metadata of each Mopseus Digital Object.

Concerning the relationships of the two models, it should address the main differences of the two models. Mopseus defines particular relationships with clear semantics, while PREMIS defines relationship categories. Based on this clarification, the relationships between a Mopseus Digital Object and its Datastreams are Structural. In particular Mopseus data model enriches the vocabulary of PREMIS structural relationships and this is obvious by the descriptions of the relationship semantics presented in Table 1. The existence of at least one Datastream that hold the main metadata of a Digital Object expressed in Dublin Core terms is mandatory for Fedora-commons repositories, including Mopseus, defining thus a dependency relationship. The modifications of the Digital Objects, generated by the performance of a Workflow or State, define derivation relationships between a Digital Object.

Since a Container is mapped to an Intellectual Entity, the relationships between a Digital Object and a Container are defined as a Relationship between different types. This type of relationships employs the PREMIS vocabulary, enriched by a variety of terms that belong to the DCTERMS Relations vocabulary. Finally the relationships between a Digital Object and a State of a Workflow as well as the relationships between the State of a Workflow and an Agent are categorized to the PREMIS categories of relationships between different types of entities. It should be noticed that these relationships are expressed similarly by both the models: The information of the domain entity of each relationship includes a pointer to the identifier of the instance of the range entity. For instance the metadata of the State of a Workflow contain the identifier of Digital Object which participates to the State.

Given the mapping of the two models, the next step is the automated generation of valid PREMIS metadata. This process is based on an XSLT document which retrieves the metadata kept in the Datastreams of a Mopseus Digital Object and writes them in an XML document that follows the PREMIS syntax; this document is stored in a new Datastream, which is updated on each modification of the Digital Object. This process is triggered when a new Digital Object is ingested in Mopseus as well as at each modification of it. This process is described in the next paragraph.

3.3. Generating PREMIS Metadata

Fedora-commons keeps a log of the operations that take place in the repository encoded in FOXML. Mopseus keeps a more detailed log in PREMIS. The service that maintains the log is invoked whenever a user (or a service) performs a write operation on the repository. This operation mainly includes the creation of Digital Objects (manually or from a migration service), the creation or modification of Datastreams, the creation or deletion of RDF relations, etc. For each Digital Object there is a log kept in the PREMIS Datastream. A sample of the log can be seen in Figure 5. Most information that is required for the creation of the PREMIS log is taken from various Digital Objects of the config namespace. For instance, the eventIdentifierType is encoded in the config:repository XML Datastreams whereas information regarding different services can be found in the config:services Digital Objects. Regarding the relationships, each Mopseus relationship described in

the config:ontologies Digital Object, is mapped to the corresponding PREMIS relation type.

```
<premis
xmlns:premis="http://www.loc.gov/standards/premis" >
   <premis:objectIdentifier>

<premis:objectIdentifierType>hdl</premis:objectIdentifierType>
<premis:objectIdentifierValue>iid:1011</premis:objectIdentifierVa
lue>
   </premis:objectIdentifier>
   <agent>
      <agentIdentifier>
         <agentIdentifierType>uid</agentIdentifierType>
         <agentIdentifierValue>13</agentIdentifierValue>
      </agentIdentifier>
      <agentName>Dimitris</agentName>
      <agentType>user</agentType>
   </agent>
   <agent>
      <agentIdentifier>
         <agentIdentifierType>servlet</agentIdentifierType>
<agentIdentifierValue>org.dcu.mopseus.DigitalObject</agentIdenti
fierValue>
      </agentIdentifier>
      <agentName>modifyObject</agentName>
      <agentType>service</agentType>
   </agent>
   <premisEvent:event
xmlns:premisEvent="http://www.loc.gov/standards/premis/v1">
      <premisEvent:eventIdentifier>
<premisEvent:eventIdentifierType>MIS</premisEvent:eventIdentif
ierType>
<premisEvent:eventIdentifierValue>modification</premisEvent:ev
entIdentifierValue>
      </premisEvent:eventIdentifier>
      <premisEvent:eventType>modify digital
object</premisEvent:eventType>
      <premisEvent:eventDateTime>2010-05-
04T17:48:39</premisEvent:eventDateTime>
      <premisEvent:eventDetail>modify digital object (dLabel=,
state=A)</premisEvent:eventDetail>
      <premisEvent:linkingAgentIdentifier>
<premisEvent:linkingAgentIdentifierType>uid</premisEvent:linkin
gAgentIdentifierType>
<premisEvent:linkingAgentIdentifierValue>13</premisEvent:linki
ngAgentIdentifierValue>
<premisEvent:linkingAgentRole>user</premisEvent:linkingAgent
Role>
      </premisEvent:linkingAgentIdentifier>
      <premisEvent:linkingAgentIdentifier>
<premisEvent:linkingAgentIdentifierType>servlet</premisEvent:li
nkingAgentIdentifierType>
<premisEvent:linkingAgentIdentifierValue>org.dcu.mopseus.Digit
alObject</premisEvent:linkingAgentIdentifierValue>
<premisEvent:linkingAgentRole>servlet</premisEvent:linkingAge
ntRole>
      </premisEvent:linkingAgentIdentifier>
   </premisEvent:event>
</premis>
```

Figure 5. A PREMIS Datastream

4. DISCUSSION

Many approaches towards building a digital repository with preservation functionalities have been implemented. We briefly present and compare them with the key features of Mopseus. Both CASPAR [8] and PLANETS [9] aim at providing a set of direction on creating practical services and tools for the purpose of long-term access and preservation, without providing an actual digital repository, and therefore are out of the scope of this discussion.

One of the most widely known and used repositories worldwide is DSpace [2] which provides an out of the box solution for grey literature management in institutional repositories. However, it doesn't address the preservation of its Digital Objects as efficiently as other repository platforms, it lacks in flexibility since it only allows flat and relatively simple metadata schemas and it limits the organization of Digital Objects by providing only a few level hierarchy of Digital Objects.

eSciDoc [14] is a powerful e-Research middleware infrastructure providing innovative services focusing on the researchers collaboration and the management of their resources. It is based on Fedora-commons repository management software on which a new data model is defined. An eSciDoc Object is represented by multiple manifestations organized in Components. Each component includes the manifestation metadata and the content itself. A single eSciDoc Object may be a composition of Fedora-commons Digital Objects correlated by whole/part or parent/child relationships. To facilitate the view of an eSciDoc Object as one entity, eSciDoc extends the Fedora-commons versioning mechanism by maintaining a datastream for each eSciDoc object that keeps track of all Fedora-commons digital objects modifications. Regarding preservation, eSciDoc incorporates JHOVE tool. Moreover the complexity of objects involved in e-Research, as well as the lack of appropriate metadata standards for their description, constitutes a barrier for the development of a concrete preservation strategy..

DAITSS [3] is a digital preservation repository application developed by the Florida Center for Library Automation and is intended to be used as a back-end to other systems, thus it has no public access interface, though it can be used in conjunction with an access system. The DAITSS system is a java application which handles all DAITSS functionality, a MySQL database to manage its archival collections and a storage back-end where DAITSS stores the information packages. DAITSS is designed to implement active preservation strategies based on format transformations including bit-level preservation, forward migration, normalization, and localization. It implements OAIS, it uses METS [10] and has a partial compliance with PREMIS. In short, the DAITSS is a functional digital repository application that

is able to perform on a large scale. There is no front-end to support the preservation functions.

The British Library's eJournal system [5] is a system for ingest, storage and preservation of digital content developed under the Digital Library System Programme, with eJournals as the first content stream. It is an implementation of OAIS, making use of the British Library's METS, PREMIS and MODS application profiles. The AIP is tied to the technical infrastructure of the British Library's preservation system, that consists of an ingest system, a metadata management component and an archival store, and is linked with the existing integrated library system (ILS). The eJournal data model contains five separate metadata AIPs, journals, issues, articles, manifestations and submissions, with each being realised by at least one METS document. Descriptive metadata are stored as a MODS extension to the METS document, while provenance and technical metadata are captured as PREMIS extensions. Events related to the digital material are being recorded as provenance metadata and can be associated with any object type. The British Library's eJournal system is an example of use of a combination of existing metadata schemas to represent eJournal Archival Information Packages in a write-once archival system.

RODA is an open source service-oriented digital repository developed by the Portuguese National Archives. RODA is based on existing standards such as OAIS, METS, EAD and PREMIS and has the Fedora Commons at the core of its framework. RODA specifies workflows for each off the three top processes of the OAIS model (ingest, administration and dissemination). Every Digital Object being stored in RODA is subjected to a normalization process. RODA makes use of the Fedora main features adding to them a set of RODA Core Services. RODA also provides a web interface to allow the end user to browse, search, access and administrate stored information, metadata, execute ingest procedures, preservation and dissemination tasks. RODA supports a set of preservation services, such as (a) file format identification, (b) recommendation of optimal migration options, (c) conversion of Digital Objects from their original formats to more up-to-date encodings, (d) quality-control assessment of the overall migration process, (e) generation of preservation metadata in PREMIS format. RODA is a complete digital repository providing functionality for all the OAIS main units and a set of preservation services developed around Fedora.

Mopseus presents most similarities and shares a similar approach with RODA since both digital repositories are Fedora based and implement the OAIS model for storing and disseminating Digital Objects. However Mopseus does not encapsulate the Datastreams in METS documents, but correlates them semantically utilizing the internal ontology RELS-INT, while the Digital Objects are correlated with the Containers via the RELS-EXT ontology. Moreover since Mopseus is focused on small and middle sized institutions can be easily installed under different platforms and requires low implementation and support expertise. One of the most powerful features of Mopseus is the flexibility in defining and mapping of metadata schemas and generating preservation metadata. These features along with the collection migration functionality render Mopseus a repository management platform adaptive to the preservation needs of several types of small and medium sized information organizations.

5. CONCLUSIONS

Mopseus is an easily configurable open source repository management system, adaptive to the digital content and needs of a variety of information organization types. It enhances Fedora-commons platform with a powerful data model providing a set of semantically rich relationships between the content and its metadata, including information concerning the provenance of them. Moreover it provides powerful functionalities for metadata schemas definition and automated preservation metadata generation, while it offers mechanisms from migrating content from other repositories. These features enable information providers to manage and preserve their digital holdings.

Among the plans for Mopseus further development is the addition of workflow wizards to the workflow engine, to guide users to define, plan and perform content management activities using friendly and usable interfaces. Mopseus does not provide a format migration mechanism due to its low cost approach. Future work includes the development of an API on which a variety of preservation planning tools such as PLATO, and format migration tools can be incorporated in the Mopseus environment. After these improvements a large scale user-based evaluation experiment will be conducted to investigate the acceptance of MOPSEUS functionalities by the user and stakeholder (libraries, museums, archives and records management services) communities.

6. REFERENCES

[1] Angelis, S., Constantopoulos, P. Gavrilis, D., Papatheodorou, C. "A Digital Library Service for the Small", *DigCCurr 2009: Procs of the 2nd Digital Curation Curriculum Symposium: Digital Curation Practice, Promise and Prospects*, 2009. http://www.ils.unc.edu/digccurr2009/

[2] Bass, M.J., Stuve, D., Tansley, R. "DSpace – a Sustainable Solution for Institutional Digital Asset Services – Spanning the Information Asset Value Chain: Ingest, Manage, Preserve, Disseminate

Functionality", Internal Reference Specification. http://www.dspace.org/technology/architecture.pdf

[3] Caplan, P. "The Florida Digital Archive and DAITSS: A model for digital preservation", *Library Hi Tech*, 28(2), 2010.

[4] CCSDS, Reference Model for an Open Archival Information System (OAIS), CCSDS 650.0-B-1, Blue Book (the full ISO standard), 2002. http://public.ccsds.org/publications/archive/650x0b1.pdf

[5] Dappert, A., Enders, M. "Using METS, PREMIS and MODS for Archiving eJournals", D-Lib Magazine, 14 (9/10), 2008. http://www.dlib.org/dlib/september08/dappert/09dappert.html

[6] Dublin Core Metadata Initiative, "DCMI Metadata Terms", http://dublincore.org/documents/dcmi-terms/

[7] Fedora Object XML (FOXML), http://www.fedora-commons.org/download/2.0/userdocs/digitalobjects/introFOXML.html

[8] Giaretta, D. "The CASPAR Approach to Digital Preservation", *The International Journal of Digital Curation* 2(1), 2007. http://www.ijdc.net/ijdc/article/view/29/32

[9] King, R., Schmidt, R., Jackson, A., Wilson, C., Steeg, F. "The Planets Interoperability Framework - An Infrastructure for Digital Preservation Actions", *ECDL2009: Procs of the 13th European Conference on Digital Libraries*, 425-428, 2009.

[10] Library of Congress, "METS - Metadata Encoding & Transmission Standard". http://www.loc.gov/standards/mets/

[11] Meghini, C., Spyratos, N. "Viewing Collections as Abstractions" DELOS Conference 2007: *Procs of the 1st International DELOS Conference*, 207-217, 2007.

[12] PREMIS Working Group, "Data dictionary for preservation metadata: final report of the PREMIS Working Group", OCLC Online Computer Library Center & Research Libraries Group, Dublin, Ohio, USA, 2005.

[13] Ramalho, J. C., Ferreira, M. "RODA: A service-oriented repository to preserve authentic digital objects", *Open Repositories*, 2009. http://redmine.keep.pt/attachments/8/OR09-0.3.pdf

[14] Razum, M., Schwichtenberg, F.,Wagner, S., Hoppe, M. "eSciDoc Infrastructure: A Fedora-Based e-Research Framework", *ECDL2009: Procs of the 13th European Conference on Digital Libraries*, 227-238, 2009.

[15] Roberts, G. "Small Libraries, Big Technology", *Computers in Libraries*, 25(3), 24-26, 2005.

ARCHIVEMATICA: USING MICRO-SERVICES AND OPEN-SOURCE SOFTWARE TO DELIVER A COMPREHENSIVE DIGITAL CURATION SOLUTION

Peter Van Garderen
Artefactual Systems
New Westminster, Canada
http://artefactual.com

ABSTRACT

Digital curation micro-services offer a light-weight alternative to preservation systems that are developed on digital repository and framework technology stacks. These are often too complex for small and medium-sized memory institutions to deploy and maintain. The Archivematica project has implemented a micro-services approach to develop an integrated suite of free and open-source tools that allows users to process digital objects from ingest to access while applying format specific preservation policies. Inspired by a call to action in a recent UNESCO Memory of the World report, the goal of the Archivematica project is to reduce the cost and technical complexity of deploying a comprehensive, interoperable digital curation solution that is compliant with standards and best practices.[1]

1. DIGITAL CURATION MICRO-SERVICES

Instead of relying on a repository interface to a digital object store, the micro-services approach uses loosely-coupled tools to provide granular and orthogonal digital curation services built around file system storage. File system technology is long-proven and extremely robust, typically outlasting the lifespan of enterprise information systems. Making the file system the focal point of micro-services operations is noteworthy as a long-term preservation strategy because it provides archivists with the option of direct, unmediated access to archival storage. This might be necessary one day because the various layers and generations of digital preservation system components are just as susceptible to the risk of technology obsolescence and incompatibility as the digital objects they are attempting to preserve.

Basing services around a basic file system store or interface (e.g. NFS, CIFS) also reduces technical complexity for development and maintenance. As noted by the University of California's Curation Center: "since each service is small and self-contained, they are collectively easier to develop, deploy, maintain, and enhance. Equally as important, since the level of investment in, and concomitantly, commitment to, any given service is small, they are more easily replaced when they have outlived their usefulness." [2]

Each service and tool integrated into the Archivematica system can be swapped for another (e.g. replacing the UUID Linux utility with the NOID application to provide the unique identifier micro-service). As a matter of fact, the entire Archivematica system is disposable from one release to the next. Release upgrades are carried out by completely deleting one disk image containing the operating system and software suite with the newer release. This is possible because Archivematica is essentially a pipeline of services, built on top of a customized Xubuntu Linux distribution, that moves digital information packages through a series of file system directories. Together these steps process digital objects from ingest through to access, leaving the Archival Information Packages (along with backups of system metadata and configuration settings) in the archival storage file system. Cached copies of Dissemination Information Packages are uploaded to a web-based access system when processing is complete. The information packages exist completely independent from the software tools. This highlights the "permanent objects, disposable systems" 3 characteristic that is a distinguishing feature of micro-service based solutions as they have come to be defined over the past year couple of years in a series of articles, architecture documentation,

[1] Bradley, K., Lei, J., Blackall, C.. Towards Open Source Archival Repository and Preservation System, 2007. http://www.unesco.org/webworld/en/mow-open-source/ (last accessed May 4, 2010)

[2] UC Curation Center / California Digital Library. UC3 Curation Foundations, Rev. 0.13 – 2010-03-25. http://www.cdlib.org/services/uc3/curation/ (last accessed May 4, 2010).

[3] Abrams, S., Cruse, P., Kunze, J., "Permanent Objects, Disposable Systems", *Proceedings of the 4th International Conference on Open Repositories*, Atlanta, U.S.A, 2009.

specifications and software tools developed at the University of California Curation Center.[4] Taken together, this substantial body of work has formed the theoretical foundation for digital curation micro-services and has established it as a legitimate alternative to repository-based digital curation systems.

2. ARCHIVEMATICA MICRO-SERVICES

While the University of California's micro-services have their origins in the need to provide support services to their campus community, Archivematica's micro-service definitions are based on a detailed use-case and workflow analysis of the OAIS functional model and the business processes of public archival institutions.[5] These were refined through proof-of-concept projects carried out in 2009 and early 2010 at the City of Vancouver Archives and the International Monetary Fund Archives.

This process led to the specification of twenty four micro-services grouped into nine OAIS workflow categories:

Category	**Micro-Service**
1. receiveSIP	verifyChecksum
2. reviewSIP	extractPackage assignIdentifier parseManifest cleanFilename
3. quarantineSIP	lockAccess virusCheck
4. appraiseSIP	identifyFormat validateFormat extractMetadata decidePreservationAction
5. prepareAIP	gatherMetadata normalizeFiles createPackage
6. reviewAIP	decideStorageAction
7. storeAIP	writePackage replicatePackage auditFixity readPackage updatePackage
8. provideDIP	uploadPackage updateMetadata
9. monitorPreservation	updatePolicy migrateFormat

Table 1. Archivematica micro-services

Each micro-service is a set of processing steps carried out on a conceptual entity that is equivalent to an OAIS information package: the Submission Information Package (SIP), the Archival Information Package (AIP) and the Dissemination Information Package (DIP).[6]

Used together, the Archivematica micro-services make it possible to fully implement the OAIS functional model, including preservation planning. As the UC Curation Center notes, "Although the scope of any given service is narrowly focused, complex curation function can nevertheless emerge from the strategic combination of individual, atomistic services." [7]

The terminology used to define the Archivematica micro-services specifications does differ from the UC Curation micro-services specifications but there is much overlap in their scope and function. Therefore, the Archivematica project would like to do more work in the coming year to align Archivematica's micro-services more closely with the UC Curation's digital curation specifications and APIs.

3. THE ARCHIVEMATICA SOFTWARE

The Archivematica system is packaged as a virtual appliance that bundles a customized Xubuntu Linux operating system with a suite of open-source software tools. Using a virtual machine application (e.g. Sun VirtualBox, VMWare Player), the Archivematica virtual appliance can be run on top of any consumer-grade hardware and operating system. The same disk image used for the virtual appliance can also be used as a bootable USB key, a Live DVD version of the system or for bare-metal installs of networked Archivematica servers and workstations. Current Archivematica development is focused on coordinating these types of networked installations to thread high-volume ingest processes over multiple nodes and thereby scale the system up to support resource-intensive production environments.

The information packages ingested by Archivematica are moved from one micro-service to the next using the Unix pipeline pattern.[8] A Unix pipeline is a well-established system design pattern wherein a set of processes are chained by their standard I/O streams, so that the output of one process feeds directly as input to the next one.[9] In Archivematica this pattern is implemented using Bash and Python scripts together with the Unix *incrond* and *flock* utilities.

[4] Curation Micro-Services. http://www.cdlib.org/services/uc3/curation/ (last accessed May 5, 2010).
[5] Micro-Services. http://archivematica.org/micro-services. (last accessed May 5, 2010).
[6] ISO 14721:2003. Space data and information transfer systems -- Open archival information system -- Reference model.
[7] UC Curation Center / California Digital Library. UC3 Curation Foundations, Rev. 0.13 – 2010-03-25. http://www.cdlib.org/services/uc3/curation/ (last accessed May 4, 2010).
[8] Abrams, S., Cruse, P., Kunze, J., Minor, D. "Curation Micro-services: A Pipeline Metaphor for Repositories", *Proceedings of the 5th International Conference on Open Repositories, Madrid, Spain, 2010.*
[9] Abrams, S., Cruse, P., Kunze, J., Minor, D. "Curation Micro-services: A Pipeline Metaphor for Repositories", *Proceedings of the 5th International Conference on Open Repositories, Madrid, Spain, 2010.*

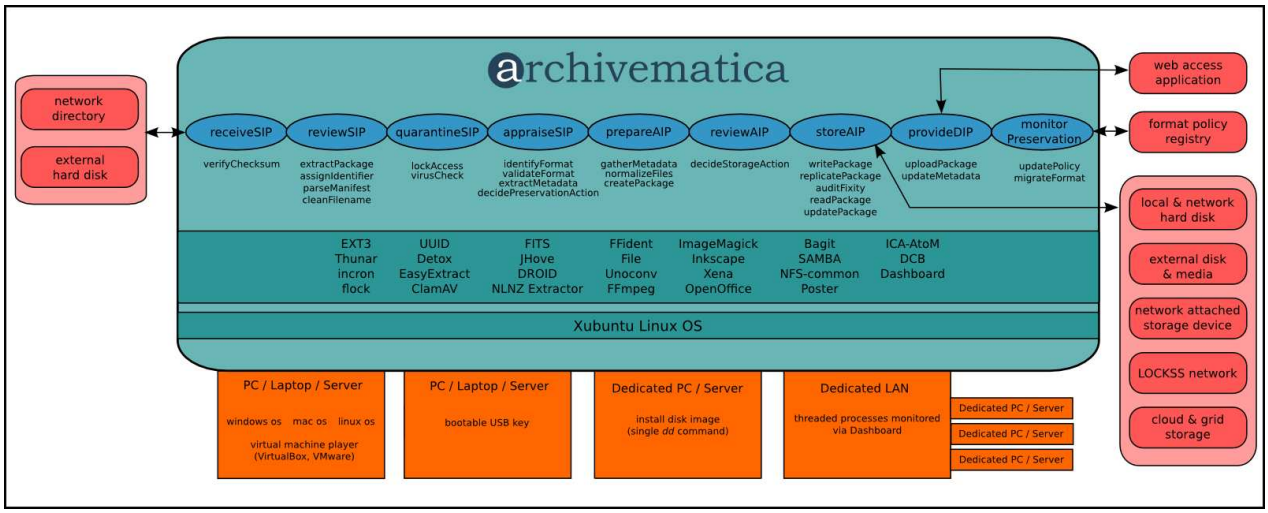


Figure 1. Archivematica architecture

Micro-service functionality is provided by one or more of the open-source software utilities and applications bundled in the Archivematica system. Where necessary, these are supplemented by Archivematica integration code written as Python scripts. Python is a proven and preferred language in large-scale integration scenarios.[10] As an interpreted language, it supports easy customization and an agile development methodology that allows for testing changes in real-time while still maintaining code integrity through the use of standard code versioning and issue tracking tools.[11]

The Archivematica 0.6-alpha release was made available in May 2010. It uses the Xubuntu desktop and Thunar file manager as its primary user interface. A web-based 'Dashboard' is currently in development to provide a more sophisticated multi-user interface that will report on the status of system events and make it simpler to control and trigger specific micro-services. The following sections describe the functionality of the current 0.6-alpha release.

3.1. Receiving files for Ingest

Archivematica provides a simple text template to create SIP manifests using qualified Dublin Core elements. However, the system will accept files for ingest with as much or as little metadata as is available. It runs the SIP through a series of ingest processing steps including unpacking, checksum verification and creation, unique identification, quarantine, format identification, format validation, metadata extraction and normalization. A variety of tools are used in each of these processes, including Easy Extract, Detox, UUID, CLAM AV, Thunar, Incron, Flock, JHOVE, DROID, NLNZ Metadata Extractor, File, FFident, File Information Tool Set (FITS), OpenOffice, Unoconv, FFmpeg, ImageMagick, and Inkscape.

3.2. Format specific preservation plans

Archivematica maintains the original format of all ingested files to support migration and emulation strategies. However, the primary preservation strategy is to normalize files to preservation and access formats upon ingest. Archivematica groups file formats into media type preservation plan (e.g. text, audio, video, raster image, vector image, etc.). Archivematica's preservation formats must all be open standards. Additionally, the choice of formats is based on community best practices, availability of free and open-source normalization tools, and an analysis of the significant characteristics for each media type. The choice of access formats is based largely on the ubiquity of web-based viewers for the file format.

Digital format identification registries (e.g PRONOM, UDFR) and conversion testbed services (e.g. Planets) are important components in a global, service-based preservation planning infrastructure. These services must be supported, in turn, by executable digital format policies that can be easily integrated into operational digital curation systems. In other words, after identifying formats, analyzing significant characteristics and evaluating risks we need to make some practical decisions about which preservation formats will ultimately be implemented and how to support these in currently operational systems using the available tools. While the analysis is important and interesting, most small and medium-sized institutions are simply asking for someone to summarize best practices and make it easy to implement them. Therefore, we need a method to document preservation format policies in a structured way to make them easy to implement in a system like Archivematica. Ideally, these policies are also publicly shared to better determine international best practices and enable risk assessment methodologies like those being

[10] See, for example, *Quotes about Python* http://www.python.org/about/quotes/ (last accessed May 5, 2010).
[11] The Archivematica Subversion repository and issue tracking list are available at Googlecode Project Hosting, http://archivematica.googlecode.com (last accessed May 5, 2010).

developed for the Preserv2 file format registry.[12] To date, it has been difficult to analyze community consensus on preservation file formats policies. These are often unreported or scattered about in reports and articles with varying degrees of accessibility.

The Archivematica project publishes its format policies and media-type preservation plans on the project wiki as these are being developed and analyzed. These will be moved to a structured, online format policy registry that brings together format identification information with significant characteristic analysis, risk assessments and normalization tool information to arrive at default preservation format and access format policies for Archivematica. The goal is to make this registry interoperable with the upcoming UDFR registry, Planets Registry and tools like the Preserv2 registry. Alternately, if these other tools are extended to support format policy implementation requirements then Archivematica could switch to use them instead. Interoperability will be facilitated by adopting the use of standards such as the eXtensible Characterization Language (XCL) specifications and by providing an RDF interface to the registry.[13] This will also facilitate the sharing of default Archivematica format policies, which might be useful to other projects and institutions.

Archivematica installations will use the registry to update their local, default policies and notify users if there has been a change in the risk status or migration options for these formats, allowing them to trigger a migration process using the available normalization tools. Users are free to determine their own preservation policies, whether based on alternate institutional policies or developed through the use of a formal preservation policy tool like Plato. The system uses a simple digital format policy XML schema that makes it easy to add new normalization tools or customize the default media-type preservation plans.

3.3. Preparing files for archival storage

Archivematica creates Archival Information Packages (AIPs) using qualified Dublin Core, PREMIS and METS elements and Library of Congress' Bagit format. As well, support for the TIPR project's Repository eXchange Package specification is currently in development.[14] Archivematica is able to interact with any number of storage systems using standard protocols (NFS, CIFS, HTTP, etc.) to allow for the flexible implementation of an archival storage and backup strategy. Standard operating system utilities can be used to provide backup functionality. Archival storage options range from local hard disk, external hard disks, network attached storage devices and LOCKSS networks (e.g. MetaArchive, COPPUL). Support for storage grids (e.g. iRODS, Bycast) and cloud storage (e.g. Amazon S3, Microsoft Azure) interfaces are also being analyzed. Ideally, the storage platform provides its own fixity check functionality (e.g. Sun ZFS, LOCKSS, iRODS) but for those that do not, a fixity check daemon will be added to Archivematica.

3.4. Making files available for access

Archivematica prepares default Dissemination Information Packages (DIP) which are based on the designated access formats for each media type. Consumers can subsequently request AIP copies but caching access copies is a much more scalable approach that will address the majority of access requests in the most performant manner, namely by reducing the bandwidth and time required to retrieve AIPs from archival storage and uploading them to the Consumer.[15] The DIP access derivatives are sent via a REST interface to a web-based application such as ICA-AtoM for further enhancement of descriptive metadata (using ISAD(G), Dublin Core, EAD, etc). These can then be arranged as accruals into existing archival descriptions to provide integrated search and browse access to the institution's analogue and digital holdings from one common web-based interface. The Archivematica Dashboard will coordinate the read and write operations of the AIP to file storage and the syncing of metadata updates between the AIPs and the access system.

4. SIMPLIFYING DIGITAL CURATION BEST PRACTICES

The project's thorough OAIS use case and process analysis has synthesized the specific, concrete steps that must be carried out to comply with the OAIS functional model from ingest to access. Archivematica assigns each of these steps to micro-services implemented by one or more of the bundled open-source tools. These, in turn, automate the use of digital curation standards (e.g. PREMIS) and best practices (e.g. Bagit). If it is not possible to automate these steps in the current Archivematica release iteration, they are incorporated and documented into a manual procedure to be carried out by the end user.

For example, in early alpha releases of the Archivematica system, some of the workflow controls (e.g. event triggering, error reporting, etc.) are handled via the Thunar file manager (e.g. drag-and-drop, desktop

[12] Tarrant, D., Hitchcock, S., Carr, L., "Where the Semantic Web and Web 2.0 Meet Format Risk Management: P2 Registry" *Proceedings of the Sixth International Conference on Preservation of Digital Objects*, San Francisco, U.S.A., 2009.

[13] The Planets XCL Project. http://planetarium.hki.uni-koeln.de/planets_cms/ (last accessed May 5, 2010).

[14] Repository eXchange Package (RXP) specification http://wiki.fcla.edu:8000/TIPR/21

[15] Wright, G., Creighton, T., Stokes, R., "FamilySearch: Extending OAIS for Large-Scale Access and Archiving" *Preservation and Archiving Special Interest Group (PASIG)*, San Francisco, U.S.A., 2009.

notifications). As the system approaches beta maturity all of the micro-services workflow will be managed and monitored via the web-based Dashboard application. Likewise, as the system matures, each service will be exposed via a command-line and/or REST API.

Focusing on the workflow steps required to complete best practice digital curation functions, instead of the technical components, helps to ensure that the entire set of preservation requirements is being carried out, even in the very early iterations of the system. In other words, the system is conceptualized as an integrated whole of technology, people and procedures, not just a set of software tools.

All software-intensive systems are dynamic, ever-evolving and, arguably, perpetually incomplete. This is particularly true for a digital curation system that must respond to changes in the technology that creates digital information, as well as the technology that is available to manage it. Therefore, the Archivematica project is a working example of the the "disposable system" concept, complemented by an agile software development model that is focused on rapid release cycles and iterative, granular updates to the requirements documentation, software code and user documentation.

5. USING THE OPEN-SOURCE MODEL TO REDUCE COSTS AND LEVERAGE KNOWLEDGE

Archivematica is still in the initial stages of development, having been made available as an alpha release earlier this year. However, by early 2011 beta versions will be implemented in production pilots at collaborating institutions. Throughout the intervening time period, the systems development will continue to be heavily influenced by the day to day feedback of its community. The Archivematica project is structured in a truly open way to encourage a grass-roots, collaborative development model which makes it easy for other institutions, projects and third-party contractors to benefit and contribute. All of the software, documentation and development infrastructure is available free of charge and released under GPL and Creative Commons licenses to give users the freedom to study, adapt and re-distribute these resources as best suits them.

No software license fees, membership dues or account registration is required for downloading Archivematica or checking out the source code from the public Subversion repository. Full documentation is provided on how to build the Archivematica virtual appliance from the source code. The community is encouraged to update the issues list and wiki pages and to join the discussion list and weekly development meetings in the online chat room.

The open-source model provides a cost-effective way to manage system maintenance expenses by freely sharing technical knowledge and documentation, providing direct access to core developers for technical support and feedback, and eliminating the need for maintenance contracts to implement release upgrades. It also encourages users to pool their technology budgets and to attract external funding to develop core application features. This means the community pays only once to have features developed, either by in-house technical staff or by third-party contractors. This new functionality can then be 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 curation technology but then cannot share that technology with their colleagues or professional communities because of expensive and restrictive software licenses imposed by the vendor.

The Archivematica project is only a year old but already the UNESCO Memory of the World Subcommittee on Technology has provided external funding to contribute to its core development, while both the City of Vancouver Archives and the International Monetary Fund Archives have sponsored the development of new features by deploying the system as part of their own internal proof-of-concept projects and contributing new code back to the project under GPL licenses.

Mature open-source communities are supported by third-party solution providers that can provide optional installation, customization, help-desk, hosting and service level agreements for those institutions that lack the capacity to implement or support their own digital curation systems. Archivematica's software development has been led thus far by Artefactual Systems, a contractor based in Vancouver, Canada that provides open-source software solutions and consulting services for archives and memory institutions. Artefactual is also the lead developer of the International Council on Archives' ICA-AtoM software project. Additional service providers are encouraged to collaborate and contribute to the ongoing development of the Archivematica platform.

6. GET INVOLVED

Like any newly launched open-source project, Archivematica is growing its network of implementation institutions, end-users, developers, solution providers, and funding sponsors. If you think that the Archivematica open-source technology, agile development methodology and micro-services conceptual framework is a good fit for your institution, then we encourage you to get involved in the project and help to define its future. You can download the application and source code or simply get started by posting questions in the discussion list, dropping in on the developers' chat room or contacting the project leads directly.

Session 5a: Preservation Planning and Evaluation

CONNECTING PRESERVATION PLANNING AND PLATO WITH DIGITAL REPOSITORY INTERFACES

Steve Hitchcock **David Tarrant** **Les Carr**

Electronics and Computer Science
University of Southampton, UK

Hannes Kulovits **Andreas Rauber**

Vienna University of Technology
Austria

ABSTRACT

An accepted digital preservation workflow is emerging in which file formats are identified and those believed to be at risk are migrated to what are perceived to be less risky formats. This raises important questions about what to convert and when, if at all. In other words, how to connect file identification and migration. This area has become known as preservation planning, and seeks to take account of a wide variety of factors that might impact preservation decisions. Broadly there are two approaches to preservation planning. One provided in some digital preservation systems is to simplify and reduce both the number of file formats stored and therefore limit the number of preservation tools needed based on accepted recommendations. A more thorough, flexible and possibly complex approach, supported by the Plato preservation planning tool developed by the Planets project, allows decisions on preservation actions to combine analysis of the characteristics of different file formats with specific local requirements, such as costs and resources. This paper shows how Plato can be integrated with digital repository software, in this case EPrints, to enable this powerful approach to be used effectively to manage content in repositories of different sizes and with varying degrees of preservation expertise and support. These tools are accessed via a common repository interface to enable repository managers, and others who do not specialise in preservation, to moderate decisions on preservation planning and to control preservation actions.

1. INTRODUCTION

Progress has been made in the development of a framework and tools for digital preservation, but so far there has been little join-up or integration of these tools to create a workflow that is accessible from within digital repositories. Typically, support for digital preservation has been aimed at national libraries and archives or enterprise-level digital libraries that might have the scope and expertise to adopt complex and costly procedures. This does not apply to all digital repositories seeking to collect and provide access to the digital outputs of research and teaching of a single institution, at universities for example, and which are now diversifying in terms of content collection and focus. This paper identifies a preservation workflow and tools that can be applied to digital repositories. We show how these tools can be accessed via a common repository interface to enable repository managers, and others who do not specialise in preservation, to moderate decisions on preservation planning and to control preservation actions.

A range of factors is driving the growth of repository content and the promise of long term preservation, and these in turn are driving the demand on the types of content a repository is expected to handle. As a result institutional digital repositories are now collecting not only peer reviewed publications and open access research but also scientific data, teaching and learning materials as well as arts and multimedia content. It is important to realise that as the range and diversification of these types of content increases, so do the problems with managing and preserving these resources. Likewise the number of tools, services and required infrastructure will also increase.

Digital preservation is now supported by a wide variety of tools, each with their own distinctive interfaces, as revealed and visualised by a series of detailed reviews of a selection of these tools by [10][11]. We are already seeing preservation tools that 'bundle' other tools to provide a specified workflow, e.g. File Information Tool Set (FITS)[1], and the emergence of preservation systems such as RODA[2] and

[1]FITS - http://code.google.com/p/fits/
[2]RODA - http://redmine.keep.pt/projects/show/roda-public

Archivematica[3], that seek to manage complexity via a single management interfaces. These aggregated tools and systems have not yet connected preservation support with the places where most new digital content is currently being deposited, stored and accessed, in the institutional repositories.

The importance of the interface in a digital system is clear from mass market consumer adoption. When launching the much publicised Apple iPad earlier this year, Apple CEO Steve Jobs said:

> "75 million people already own iPod Touches and iPhones. That's all people who already know how to use the iPad."[4]

Familiar and successful interfaces reduce barriers to entry for systems and devices and enable users to make faster progress and become more productive. For digital resources, many Web based repositories have been built on widely-used open source software such as DSpace, EPrints and Fedora. These have a common, and often underestimated, resource: their interfaces. In fact, repository software is essentially a series of interfaces for deposit, search, browse and management tasks performed by content authors and contributors, users of the information they provide, repository administrators and third-party service providers. It seems likely that additional repository services, such as preservation, should be provided through the familiar repository interface, rather than through the native interfaces of a disparate set of tools.

Although nascent repository policies don't yet state it explicitly, it is unlikely that repositories which grow on the basis of institutional requirements can escape the consequent expectation of effective content management over timescales consistent with the institutions planning horizons.

The JISC KeepIt repository preservation project is working with four specific repositories, chosen for the variety of content types which these repositories hold, to deploy an exemplar toolkit capable of helping and performing digital preservation on these repositories. By integrating a set of tools and services into existing repository software interfaces, we demonstrate not only the value to repository managers but also how taking this approach lowers the barrier to understanding and applying digital preservation.

In this paper we highlight how preservation workflow, and one particular part of that workflow, preservation planning, has been integrated within a repository interface.

[3]Archivematica - http://archivematica.org/wiki/index.php?title=Main_Page

[4]Steve Jobs, launching the iPad, January 27, 2010 http://news.cnet.com/8301-31021_3-10440943-260.html

2. PRESERVATION WORKFLOW AND REPOSITORIES

Digital libraries have long acknowledged that preservation is a vital part of the role of a repository. However, preservation is often sidelined due to the practical constraints of running a repository. Dealing with institutional-scale ingests and quality assurance with minimal staff and investment rarely leaves sufficient capacity for engaging with a preservation agenda when the creation of a concrete plan for preserving an institution's collection of digital objects may require the detailed evaluation of possible preservation solutions against clearly defined and measurable criteria.

Digital preservation is the process of storing and managing content for the purpose of continued access through changes in the technology framework over time, both to present the essential content or data (the digital bits) and, ideally, to be able to continue to represent the author.s original intent and meaning through other features. Broadly, preservation has been modeled as a set of administrative processes allied to more technical processes for digital content management and storage. Underlying the latter are the computing applications and platforms that are used to create, distribute and access digital content, now including repositories, the Web, and so on. This analysis of the purpose and practice of digital preservation has produced a consensus on a practical preservation workflow that, while it may differ in terminology, has a common core that can be represented with respect to digital objects and their formats as follows:

identification - characterisation - risk assessment - planning – action

The first and last elements of this workflow, covering actions such as format conversion, or migration, to safer formats, are the simplest to understand and tools are available to implement these processes. The key requirement now is joining these two end-processes through the more difficult, and subjective, steps of risk assessment and planning, to determine whether, and when, a preservation action such as a migration should apply.

To implement this workflow for EPrints digital repositories, KeepIt and its predecessor JISC Preserv projects have been applying tools for preservation workflow produced by some of the constituent partners in the European-wide Planets project[5] such as the National Archives (of the UK, TNA), the British Library and the Vienna University of Technology (TU Wien):

- Format identification: DROID (TNA)

[5]Planets, Preservation and Long-term Access through Networked Services http://www.planets-project.eu/

- Characterisation: XCL (XCDL, XCEL), a means of recording the significant characteristics of a digital object in an XML-based format
- Risk assessment, planning: PRONOM (TNA), PLATO (TU Wien)

KeepIt is building these tools into EPrints software through a series of plug-ins that provide access to the tools. EPrints has offered this modular application architecture since version 3.0 in 2007, and the latest version 3.2 is required to access these preservation plugins. The processes and services these tools provide can be accessed via a common interface that allows repository managers to moderate format risk assessment, and set parameters the repository software can use to make decisions on taking preservation actions such as migrating formats. Two successful workshops have been conducted to present first-hand experience of these tools and interfaces to repository managers [5][7].

The project has also contributed to the workflow by creating a format risk registry to show how format risk can become more open based on linked data principles [12]. The aim is to integrate the ability to process all preservation-related information within the repository. This includes extending to new means of representing provenance, such as the Open Provenance Model [9], and is all handled by tools that can pass XML-based information.

3. IDENTIFICATION & CHARECTERISATION

The first stage in the preservation workflow, identification, has a number of established tools. Using PRONOM-DROID[4], for example, the first Preserv project [3] created a central format profile service for repositories (available via roar.eprints.org), which revealed a heterogeneous range of formats were in use.

In Preserv2 we realised this service needs to be bought within the repository to play a useful part in a 'smarter' repository preservation workflow [6]. Repositories may not be able to reveal all stored content, and a service such as PRONOM-ROAR cannot provide information on files which are not publicly available. Additionally, some repositories may expose content incorrectly to Web harvesters such as PRONOM-ROAR - a typical example would be a PDF document that requires authorisation to access, would return an HTML page but without the required error code (HTTP 402), thus not making it clear to the harvester that this isn't the resource it requested. Managing this service from the repository administration screen provides more detailed and trusted results. Figure 1 shows this screen as it stood at the end of the Preserv2 project.

What this approach demonstrated is how repository software can work with different preservation tools, in this case for format identification, and can import, process, display and export preservation-related data in XML-based formats, and this is the basis of the latest work to integrate preservation planning with the repository.

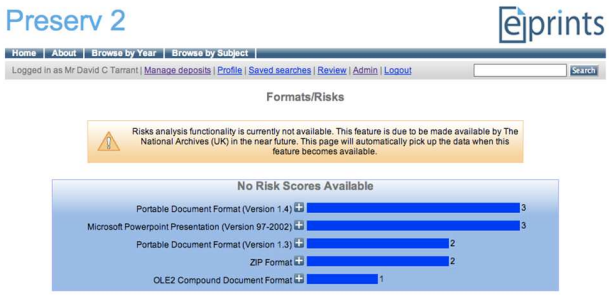


Figure 1. EPrints: Preservation interface showing file classifations

Figure 1 shows early developments on the preservation workflow in EPrints and this screen forms the basis of the rest of the work presented in this paper. At each stage of the preservation workflow this screen has been updated to become the central preservation control interface in the repository.

The identification of a file format can obscure as much as it can reveal of the essence of a digital object if we simply rely on this process to name the format. The format will have been created as the result of using one or more software applications, which will have allowed the creator to embed certain required features in the object, and also allows the user to recreate these features. Given the power of modern applications, it is possible that creator and user, or other interested stakeholders such as archivists, may seek to exploit different features in the object. This characterisation of digital objects, according to the viewpoint of different stakeholders and the significance they may attach to features, is an emerging area of interest in digital preservation and presents additional tasks in the preservation workflow [8].

4. RISK ANALYSIS

While the preservation workflow has become clear, the basis for making decisions on how to implement each stage of the workflow has not because in many cases a detailed risk analysis is not available. What risks are posed by a given file format, by an alternative format and by the tools used for conversion, and how can these risks be quantified? To some extent these questions can be answered by registries such as PRONOM. As part of the Preserv2 project the National Archives (UK) constructed a series of risk categories and also a schema for assigning a risk value to each of these categories. The idea is to enable registries to generate numerical scores for format risk.

Using these categories and a hypothetical scoring system, the preservation interface in EPrints was enabled to obtain these risk scores from PRONOM and then

display these in a traffic light scale depending on the score returned. Figure 2 shows the same preservation screen in EPrints, this time displaying information on format risk. Note that risk information depicted here is for demonstration purposes only and should not be considered an indication of actual format risk.

Figure 2. EPrints: Preservation interface with risk score data

Quantitative research in this area is still fairly new, so it remains more difficult to provide dynamic risk information as opposed to the comparatively static information describing file format. For example, the number of available tools and software products which can read a particular format, which are relevant factors in a risk analysis for that format, is likely to change more regularly than the documentation and encoding of the format itself (which typically only changes with a new version of the format).

[12] looked at possibilities to crowd source such information from publishers of linked data [2] and demonstrates the immediate benefit of combining information from PRONOM and dbpedia (wikipedia) and, potentially, other sources of format information. Ongoing work in this area is being performed by the Unified Digital Formats Registry (UDFR) project.

5. PRESERVATION PLANNING

By this stage in the preservation workflow we have identified a digital object by its format, we have analysed the characteristics we consider to be significant and that might impact preservation decisions on the object, and we have begun to quantify the risk associated with different formats and actions. All of this helps us to connect preservation action decisions with specific digital objects. Now we have to consider how this process might scale for the growing content of a digital repository. Digital content has been shown to grow at great rates, e.g. Planets market survey, so a way is needed to plan, record and, when necessary, update the whole analysis, and then to apply the plan when and where required to automate the outcomes. This is the preservation planning stage of the workflow.

Preservation planning involves evaluating available solutions against clearly defined and measurable criteria and arriving at a concrete plan of action based on these evaluation results. It is important to realise that this is both a static process which should be undertaken at the beginning of any project (alongside risk analysis) and also reactively as risks arise during the project itself.

Both Planets and EPrints believe the most effective form of file format preservation is done reactively. This implies that risk analysis is a constant process which identifies specific cases where a preservation plan may need to be constructed. This also avoids the risk of a preservation plan becoming outdated at a later stage, potentially causing a greater problem.

The remainder of this paper focuses on this small part of the preservation workflow, in effect the preservation planning sub-workflow, shown below, and considers enhancements to the repository and preservation planning tool to enable these to work together to provide a fully capable preservation solution.

collection gathering - planning - management - action - review

6. COLLECTION GATHERING

A preservation plan relates to a set of digital files. This is the first role of the repository, enabling the user to identify the set of files at risk, using the interface described earlier, and to select a collection of these files ready to import into the planning tool.

Figure 3. EPrints: Preservation Actions Panel

Figure 3 shows the Preservation Actions panel within EPrints. This panel corresponds to a single format that is identified as being at risk. Inspection shows this interface is also used in later stages of the planning process. Using this panel the user can choose the quantity of files required from the risk category, which can then be download and imported into the Plato planning tool.

To ensure the preservation plan is robust to the significant characteristics of a file format, it is important to provide the planning tool with a selection of files exhibiting as many of these characteristics as possible. While this can be done by performing deep analysis on the files, initially it was decided to settle on the following simple set of criteria, applied by EPrints, to select the files:

```
If more than 1 file requested:
        Provide Newest and Oldest
If more than 3 files requested:
        Also provide Largest and Smallest
Then
        Provide a random selection
```

Although simple, this approach should reveal a variety of factors. It is envisaged that further tools could be built into the repository to examine files in more detail, to provide a greater selection of files containing a wider range of significant characteristics. However, this should be done objectively such that the characteristics can be clearly identified, ensuring users understand why each file has been selected. A loss of understanding about why each file has been chosen could potentially be more wasteful than beneficial. This is certainly an area for further investigation and with the proliferation of tools in this area should certainly be one of the easier targets to achieve.

7. FORMING A PRESERVATION PLAN

With a collection of files gathered and imported to the planning tool, the next stage follows the workflow set by Plato, which is designed to guide the user through a set of experiments to make decisions and to formulate a preservation plan. The basic Plato workflow has four stages:

define requirements - evaluate alternatives (run experiments) - analyse results - build and validate preservation plan

Plato encourages users to think carefully about their decisions and allows improvements to be made iteratively at any stage of the workflow to obtain the best result. Each of these stages can be replayed at any time if it is realised that data is missing, but the eventual target remains the choice of one of the alternatives and making it the basis of the preservation action plan.

While there are many valid preservation actions, including do nothing, hardware/software emulation and migration, in the context of this report we are only considering migration. While "do nothing" is already supported as a no action plan, emulation with a repository context may require a substantial amount of further work to achieve.

With a great number of tools available to migrate one format to another, the choice of which new format to use can be quite complex, thus the need for a well defined set of requirements becomes ever more important.

Plato [1] defines three main stages in the preservation planning workflow:

1. Requirements definition: The important first step that defines the later evaluation criteria. Requirements from a wide range of stakeholders and influence factors have to be considered for a given institutional setting. This involves curators and domain experts as well as IT administrators and consumers. Requirements are specified in a quantifiable manner, starting with high-level objectives and breaking these down into measurable criteria, thus creating an objective evaluation tree. The requirements stage is also used to specify the significant characteristics and sustainability requirements which any plan must fulfill.
2. Evaluation of potential strategies: a series of tools are picked which suit the requirements outlined. Each file in the test set is migrated, and each successfully migrated file (e.g. an output was obtained) is stored for evaluation.
3. Analysis of the results: the results of the test migrations are evaluated against the requirements. As the requirements are weighted, this allows the planner to produce a well-informed recommendation for a preservation solution.

Finally, Plato allows a preservation plan to be exported. Although this could be the final plan, a plan can be exported at any time during the workflow and includes details of all requirements, tools selected as well in-line encoded copies of the files that were uploaded as the test set. The resulting action plan constitutes a small part of the total preservation plan, which is represented in XML and designed to contain a full record of all decisions and criteria that were evaluated in Plato. This plan can be reloaded into Plato at a later date for re-evaluation to ensure the defined action is still the most applicable.

Once a final preservation action is selected, this data is then also exported as part of the XML preservation plan, this small part of the plan conforms to a schema which has been specifically designed by the EPrints and Planets collaboration to be parsed easily by third party tools. This ability to output a clear and concise action plan that can be interpreted by other tools is a critical feature of Plato. A sample action plan is depicted by figure 4.

```
<ActionPlan action="migrate">
<tool>
  <toolIdentifier  uri="http://dbpedia.org/data/ImageMagick"  version="6.5.1-
0"/>
  <parameters>
    -verbose -compress None -quality 100 %INFILE% %OUTFILE%
  </parameters>
  <targetFormat mimetype="image/png" extension="png"/>
</tool>
</ActionPlan>
```

Figure 4. Plato: Example action plan

Figure 4 represents the minimal amount of information a repository such as EPrints needs to understand and to apply the action plan. Outlined below are the main elements of this plan and the reasons for their inclusion:

ActionPlan key:action values:migrate,emulate,none

Broadly outlines the plan and the chosen strategy. Note, it is important not to assume just because the preservation plan does not contain an ActionPlan section that the preservation action is to do nothing. An action plan which does nothing should state this implicitly.

ToolIdentifier key:uri values:Semantic URIs

Identifies, using a globally unique identifier, the tool used in the action plan. In our example this tool is ImageMagick.

ToolIdentifier key:version values:Version Number

This is a vital field if you wish to verify the translation achieved by the tool is exactly the same as the one performed by the evaluation in Plato. A different software version here is likely to generate a slightly different file as it will write its own version information into the file.

Parameters

This field defines the parameters used in the execution of the tool. For the purposes of simplicity we have chosen the constants %INFILE% and %OUTFILE% to represent how files are parsed to the tool.

TargetFormat keys:mimetype,extension

The mimetype and extension define the resultant format of the migration. Although only one should be required, some mimetypes can have multiple extensions which each tool may choose to use, thus it is handy to include both.

Plato thus forms one part of the preservation workflow which takes a set of inputs, in this case a set of files of a single format which have been identified as being at risk, to help the user produce a preservation plan, the output. The key to making these components work in a repository environment is to be able to handle both the inputs to the planning process and the subsequent output.

8. PRESERVATION PLAN MANAGEMENT

Once the preservation plan has been formed, we now require the repository to accept this plan to be managed, preserved and acted upon. As shown in figure 3, EPrints allows the preservation plan to be uploaded directly via the same preservation actions user interface used to download the original at-risk repository files to Plato. Each file format can be related to a single preservation plan that can be uploaded via the preservation interface, and these can be managed using the screen section shown in figure 5. This shows a preservation plan has been defined for the GIF image format from 1987, and that this plan, uploaded in March 2010, has performed an action on a single file in the repository, defined by the action plan outlined in the previous section.

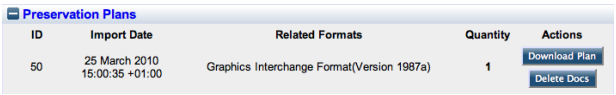


Figure 5. EPrints: Preservation plan management panel

This successful migration means that EPrints has been able to find the tool defined by Plato and act upon the plan in order to migrate not only the files that were given to Plato in the test data set but also all other files of this type in the repository. By handing back the task of mass migration of all files of the identified type to the repository alleviates the scalability issues. Here the repository is already handling this quantity of content and may as a result already be linked to services and service providers who can assist with on-demand processing for larger operations such as migration. Similar concepts for mass identification, that can be mapped to mass migration, are discussed in [6].

9. VALIDATING RESULTS

EPrints displays preservation results for the administrator via a screen that tracks file formats, showing the quantity of files in each format classified according to the associated risks on a traffic light scale, where red represents high risk and green low risk. This screen also presents formats resulting from actions performed by the preservation plan. Figure 6 is a snippet from this interface showing a set of high risk (GIF 1987a) files in the low risk category, reflecting the new risk category for the migrated files, with the red bar reflecting which category these files would be in had the migration not taken place.

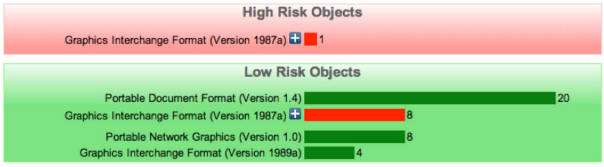


Figure 6. EPrints: Preservation interface showing migrate files

A single file of the same format remains in the high risk category because it has no migrated version yet. This could be a newly uploaded file or one which failed the migration. Pressing the '+' button would allow the plan to be manually executed on this file, and also allows the repository administrator to see when the file was created and provide other information about recent related processes. Note that EPrints never deletes any files, i.e. the originals, without explicitly being asked to do so by a depositor, editor or repository manager.

EPrints also updates the record, or abstract page relating to each item, to show that a preservation action has taken place. Figure 7 shows a single file, in PNG

format, that has been migrated from the original GIF version. Both the migrated and original files are shown in this instance with the relation between them clearly displayed. This also demonstrates part of the provenance information stored within EPrints relating any migrated files not only to the originals but also to the preservation plan that caused that migration.

Figure 7. EPrints: Abstract screen for migrated record

Figure 7 shows a public-facing EPrints abstract page containing details of a migrated file which is the result of a preservation plan. The original, pre-migration, file is kept but simply has less prominence.

10. PROVENANCE AND PRESERVATION METADATA

Provenance is an important aspect of digital preservation to establish the authenticity of objects. By migrating an object the repository is creating a new version which, for it to be authentic in the eyes of the user, needs a full set of preservation metadata detailing why and how this new version was created. We have shown how the repository stores the full preservation plan containing all the requirements and decisions made during the planning process. It should also be possible to find which files resulted from that plan or, vice-versa, which plan a file was a result of.

EPrints uses the Open Provence Model (OPM), and data stored by EPrints relating to preservation and migration can be easily serialised according to this model. OPM [9] defines a minimal set of core elements, including the following which are detailed in terms of their application in EPrints:

- WasDerivedFrom: relates the original to the migration file. In EPrints this is the two-way relation isMigrationVersionOf and hasMigratedVersion.
- WasInformedBy: relates the preservation plan to the action which took place. In EPrints this is a one-way relation defined by isMigrationResultOf.
- WasGeneratedBy: holds many important roles, both to define which tool generated the plan, but also which tool was used to perform the migration. In EPrints this information is left in the preservation plan from where it can be sourced.

With EPrints and many other repository platforms accepting arbitrary linked data (triples) relations between objects, adding this type of data to an existing record is well supported. Both the objects and the preservation plan (which is also an object) obtain a persistent identifier, which can be used to relate the objects.

Figure 8 provides an overview of the relations between the objects in EPrints that are part of this preservation process, including the original and migrated files as well as the preservation plan and file format which relates the original file to the preservation plan. This shows some of the key actors involved in this plan and the relations between them. For clarity, this omits the preservation actions related to the migrated file via the preservation plan.

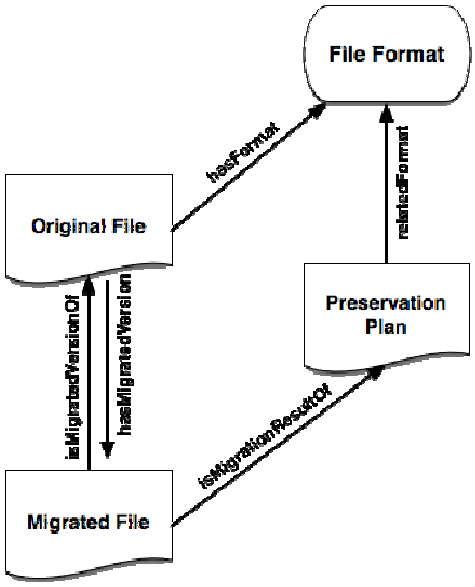


Figure 8. Provenance and preservation metadata structure

11. CONCLUSION

A preservation workflow has become established to manage the file formats of digital objects and take preemptive actions to ensure the objects remain accessible and usable as originally intended. Critical new developments have been described in this paper that enable this workflow to be managed from sources such as digital repositories that are seeing rapidly increasing volumes of content deposited yet are often managed with few resources for preservation. Key to implementing the workflow within these repositories is to use and adapt familiar repository interfaces rather than require administrators to learn new interfaces for many preservation tools required to implement the workflow. We have shown how the results from a powerful preservation planning tool, Plato, can be applied and controlled using this repository preservation interface. Importantly, through a series of workshops, we have also shown that by making these tools look familiar to the repository managers, that the barrier to learning the issues with digital preservation and understanding the responsibilities is lowered. Subsequently through the KeepIt project, these tools

have now been rolled out to a number of partner institutions as well as being made available freely online. While it is clear there is still work to be done in some areas, completing the join up of the preservation workflow from characterisation to preservation action within a familiar interface represents a huge leap forward for digital preservation.

12. REFERENCES

[1] Becker, C., et al., ”Systematic planning for digital preservation: Evaluating potential strategies and building preservation plans” *International Journal on Digital Libraries (IJDL)* December 2009 http://www.ifs.tuwien.ac.at/ becker/pubs/becker-ijdl2009.pdf

[2] Berners-Lee, T., ”Linked Data” *W3C Design Issues* 2006

[3] Brody, T., et al., ”PRONOM-ROAR: Adding Format Profiles to a Repository Registry to Inform Preservation Services”, *International Journal of Digital Curation*, Vol. 2, No. 2, December 2007 http://www.ijdc.net/ijdc/article/view/53

[4] Brown, A., ”Automatic format identification using PRONOM and DROID” *The National Archives, Digital Preservation Technical Paper* 2005

[5] Field, A., et al., ”Digital Preservation: Logical and bit-stream preservation using Plato, EPrints and the Cloud.” *In 13th European Conference on Digital Libraries (ECDL)*, 27 September 2009, Corfu http://eprints.ecs.soton.ac.uk/17962/

[6] Hitchcock, S., et al., ”Towards smart storage for repository preservation services” *In: iPRES 2008: The Fifth International Conference on Preservation of Digital Objects* 29-30 September 2008, London, UK http://eprints.ecs.soton.ac.uk/16785/

[7] Hitchcock, S., et al., ”Digital Preservation Tools for Repository Managers 4: Putting storage, format management and preservation planning in the repository” *In: KeepIt course module 4*, 18-19 March 2010, Southampton, UK http://eprints.ecs.soton.ac.uk/21029/

[8] Hitchcock, S., et al., ”Digital Preservation Tools for Repository Managers 3: Describing content for preservation” *In: KeepIt course module 3*, 2 March 2010, London, UK http://eprints.ecs.soton.ac.uk/21001/

[9] Moreau, L., ”The Open Provenance Model Core Specification (v1.1)” 21 Dec 2009 http://eprints.ecs.soton.ac.uk/18332/

[10] Prom, C., ”Archive for category Software” *Practical E-Records*, Carious Entries 2009-10, http://e-records.chrisprom.com/? cat=3

[11] Prom, C., ”PLATO (Digital Preservation Planning) Software Review” *Practical E-Records*, April 25, 2010 http://e-records.chrisprom.com/? p=1082

[12] Tarrant, D., et al., ”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*, October 5-6, 2009, San Francisco http://eprints.ecs.soton.ac.uk/17556/

FULL PAPER

EVALUATION OF BIT PRESERVATION STRATEGIES

Eld Zierau
The Royal Library of Denmark
Dep. of Digital Preservation
P.O.BOX 2149
1016 Copenhagen K
Denmark

Ulla Bøgvad Kejser
The Royal Library of Denmark
Dep. of Digital Preservation
P.O.BOX 2149
1016 Copenhagen K
Denmark

Hannes Kulovits
Vienna University of Technology
Inst. of SW Tech. & Interactive Systems
Favoritenstraße 9-11/188/2
1040 Vienna
Austria

ABSTRACT

This article describes a methodology which supports evaluation of bit preservation strategies for different digital materials. This includes evaluation of alternative bit preservation solutions. The methodology presented uses the preservation planning tool Plato for evaluations, and a BR-ReMS prototype to calculate measures for how well bit preservation requirements are met.

Planning storage of different types of data as part of preservation planning involves classification of the data with regard to requirements on confidentiality, bit safety, availability and costs. Selection of storage parameters is quite complex since e.g. more copies of data means better bit safety, but higher cost and higher risk of compromising confidentiality.

Based on a case study of a bit repository offering differentiated bit preservation solutions, the article will present results of using the methodology to make plans and choices of alternatives for different digital material with different requirements for bit integrity and confidentiality. This study shows that the methodology, including the tools used, is suitable for this purpose.

1. INTRODUCTION

This paper explores how bit preservation strategies can be evaluated against different bit repository solutions. A preservation strategy presents the chosen solution for bit preservation. The bit preservation strategy must ensure that the actual bits remain intact and accessible at all times, and is the starting point for further preservation actions. Functional (logical) preservation, which assures that the data remains understandable through further preservation actions are *not* part of bit preservation.

The research question we want to investigate is how we can evaluate requirements for a bit repository. This concerns e.g. bit safety, confidentiality and cost for alternative bit preservation solutions.

Requirements for bit preservation can be hard to express on the general level. As Rosenthal et al. notes it is a question of risk analysis [5]. We will in this article take an approach where requirements are defined in terms of importance of risk preventions. Formulation of the requirements is primarily based on the ISO 27000 series [2], complimented with analysis of bit safety [4], and own experiences.

Bit preservation implementation is hard in itself, and a lot of the technical and organisation details on the final bit preservation solution can be crucial for how well it fulfils requirements for risk prevention as explained in [6]. The challenge here is to express how different combinations of ways to store and check data copies will meet requirements.

The article presents a methodology which can help in evaluation of bit preservation strategies against choice of bit preservation alternatives. The methodology seeks to separate evaluation of requirements from the complexity of bit preservation in order to make an evaluation more clear and understandable. This is done using a tool which we call: Bit Repository – Requirement Measuring System (BR-ReMS). It is a prototype, which contains the details separated from the requirements. The BR-ReMS results are scores on how well a bit preservation solution prevents different risks.

The methodology uses the preservation planning tool Plato to evaluate how well potential bit preservation strategies meet bit preservation requirements (as a result of the BR-ReMS). Plato is a Planets tool for specification of preservation plans, primarily on logical preservation strategies [1]. In this article we will use it for evaluation of bit preservation strategies only.

In order to investigate the soundness of the methodology, we include three cases of digital material with different requirements for confidentiality and bit safety.

2. METHODOLOGY

The methodology behind our evaluation of bit preservation strategies is based on assumptions on how we can express bit preservation strategies and include requirements, assumptions on parts in bit preservation solutions, and which tools we use for the evaluation.

2.1. Assumptions on Bit Preservation Strategies

We will assume that we can evaluate a bit preservation strategy in terms of evaluating requirements against solutions. This conforms to the definition of requirements that document constraints and influence factors on potential preservation strategies in Plato.

In our case study, the bit repository requirements are assumed to be best formulated in terms of risk prevention. There are many other ways to formulate the requirements, for example at a much more detailed technical and organisation level. It should be noted that the methodology would also apply if another approach were chosen for requirements. The change would only have to be made in the set-up of the BR-ReMS and Plato tools used.

2.2. Assumptions on Bit Preservation Solutions

We will assume that bit preservation solutions can be represented as a solution offered by a conceptual bit repository (BR). A BR is a repository with a technical system managed within organisations with all aspects of an OAIS[1] system as defined in [6].

We will need to make assumptions on how bits are preserved. The assumption is that data must be kept in more copies represented as replicas. Each replica is a copy of the data stored in a pillar. A pillar is defined as a unit, which can be seen and analysed as an individual unit at the abstract level.

Replicas located on different pillars must be coordinated and possibly checked at a general BR system level. This architecture is illustrated in Figure 1. The assumption is only made on the conceptual level. This means that this architecture applies for a Danish National BR under implementation [6], or on a LOCKSS[2] system, or a SAN[3] system with backup.

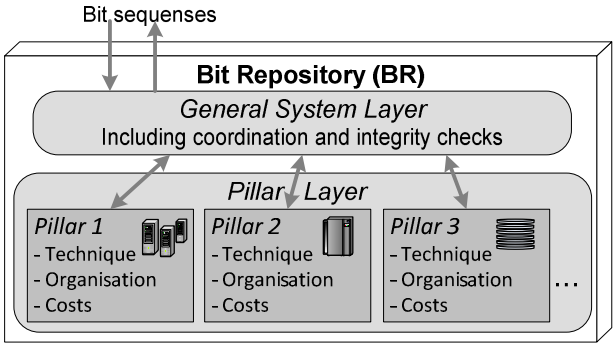


Figure 1. Bit repository with pillars.

Each pillar has different characteristics e.g. the type of media on the technical side, the physical location and procedures for operation on the organisational side, and the costs of using a pillar as basis for a replica. Similarly the general system layer has different characteristics e.g. communication protocol, speed, and bit audit frequency.

For simplicity we assume that bit integrity checks are made on a voting system based on checksums. For example, three replicas participate in a voting, where two replicas agree on a checksum, but the third does not. In this case the third replica will be reported as the faulty one. Voting is based on checksums instead of full comparisons for efficiency reasons.

An additional assumption is that a replica can be a derived replica in form of a checksum. We will call this a checksum replica instead of a full replica which contains a full copy of the data. Checksum replicas are included, since choice of having checksum replicas can increase bit safety at a low cost, but the risk analysis will e.g. depend on its physical location. This is based on Danish experiences explained in [6].

2.3. Using the BR-ReMS and Plato

At the start of this study we intended only to use the Plato tool for evaluation of bit preservation evaluation. However, it quickly became obvious that the specification of a bit preservation strategy and the influence of changing a single characteristic on a pillar were too complex to express directly in Plato.

This lead to the development of the BR-ReMS prototype, which is used to encapsulate the details on different characteristics for parts of the BR, and how they in combination change the measured levels of e.g. bit safety and confidentiality risks. The BR-ReMS produces the results which can be used in evaluation of a bit preservation strategy defined in Plato. This is illustrated in Figure 2.

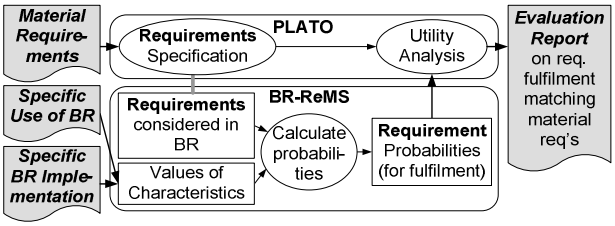


Figure 2. BR-ReMS and Plato.

The white square represents specified data whereas the grey squares represent actual input and output. The circles are processes where the arrows give directions of the information flow. The thick grey line indicates that requirements considered are the same.

3. SETUP OF REQUIREMENTS AND TOOLS

In order to understand how the methodology works, we here give a description of the set-up of the tools, as well as the choices made in definition of the requirements.

[1] OAIS (Open Archival Information System). 2002. ISO 14721:2003.
[2] See http://lockss.stanford.edu/lockss/Home
[3] See http://en.wikipedia.org/wiki/Storage_area_network

3.1. Plato

Plato is a preservation planning tool developed within the Planets[4] project and is available to the public in an open source version[5]. It has been developed in order to provide a systematic approach for evaluating potential alternatives for preservation actions and building thoroughly defined, accountable preservation plans for keeping digital content alive over time. The method follows a variation of utility analysis to support multi-criteria decision-making procedures in digital preservation planning. The selection procedure leads to well-documented and transparent decisions.

The applicability and usefulness of the tool has been validated in a series of case studies involving different organisations and digital content such as described in [3]. However, instead of evaluating migration tool with respect to the requirements, we here use the approach to analyse the results of the BR-ReMS for alternative bit preservation solutions. The results of the BR-ReMS are analysed and aggregated, corresponding to evaluation of the bit preservation strategy. Further details on this process can be found in [1,3].

3.2. The BR-ReMS Prototype

The BR-ReMS prototype is developed using Microsoft Access 2003. The set-up for specific cases is based on requirement definitions and definitions of different characteristics. A requirement definition includes definition of a function which calculates to which degree the requirement is met for different BR solutions. The calculations are based on the specified characteristics. This is exemplified in Figure 3, where the boxes with dashed lines are templates, and their use is indicated by thick grey lines.

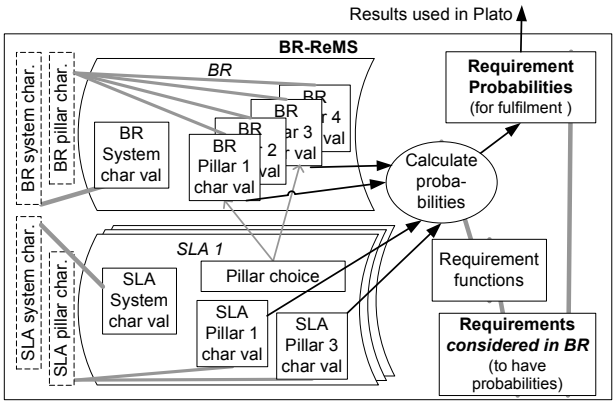


Figure 3. The BR-ReMS prototype.

As illustrated in Figure 3 there are different types of characteristics. There are the BR characteristics which are predefined by the actual 'BR implementation' (see Figure 2). And there are the service level agreement (SLA) characteristics, which are defined by individual SLAs for 'specific use of the BR' (see Figure 2). A SLA is defined as the agreement of level of service between the unit responsible for the BR and a user preserving bits in the BR, e.g. on which pillars the replicas are placed, and for each pillar, whether it is checksum or full replica. Note that we only talk about a conceptual SLA for a conceptual BR, i.e. there are no requirements to degree of formality and whether the SLA is involves several organisations operating different parts of the BR.

The BR characteristics are divided into BR general system characteristics (e.g. for transmission of data or coordination ensuring hardware/media migrations are not performed at the same time), and BR pillar characteristics for the individual pillars (e.g. hardware type, or characteristics related to natural disasters). In the same way the SLA characteristics are divided into SLA general system characteristics (e.g. bit audit frequency) and SLA pillar characteristics (e.g. digital objects are checksum replicas or full replicas).

The characteristics are defined in two steps. Firstly, the characteristic itself is defined. Secondly, the value(s) of the characteristic are defined for the different parts of the BR and individual SLAs.

Requirements are defined along with their functions. These functions can be quite complex and depend on different types of characteristics. In order to ease the calculation general functions are introduced for each pillar characteristic (both BR and SLA pillar characteristic) to be calculated across the pillars selected in a SLA. Some sub-functions also go across pillar characteristics and general system characteristics, as for example comparing frequency of bit audits with Mean-Time-To-Failure on the different media. For such purposes intermediate result characteristics are introduced which can be used in more complex calculations. Note that calculation over more pillars will work differently depending on the requirement it belongs to. For example, in calculation of bit safety requirements, adding a replica will always lower the risk of loosing bits. On the other hand in calculation of confidentiality requirements, the general rule is that adding an extra full replica will mean higher risk for lack of confidentiality.

The setup of the functions is still on a prototype level at this stage. The functions could be better described and tuned by use of more complex calculations e.g. using statistically models for error occurrence etc.

Since the details on calculation of how requirements are met are important, the BR-ReMS also offers reporting on definition of values of characteristics and definition of function used. Such reports would be input for a thorough evaluation of a bit preservation strategy or to audit actual implementation of BR parts.

3.3. Requirements used in Plato and in BR-ReMS

The definition of the requirements represented in the SLA will express the bit preservation strategy to be

[4] Preservation and Long-term Access through NETworked Services (Planets). See http://www.planets-project.eu/

[5] See http://www.ifs.tuwien.ac.at/dp/plato

evaluated in Plato, as well as the requirements that the BR-ReMS produces results for. That means the requirements must be specified in both the BR-ReMS and Plato.

We will here base requirements on the ISO 27000 series [2], as far as possible. The reason for this choice is that the ISO standard is a commonly used standard in repositories. It includes confidentiality (ensuring that information is accessible only to those authorised to have access) and integrity (safeguarding the accuracy and completeness of information and processing methods) as some of the main risk areas for information security. These are also the aspects that we have chosen to focus on in this article. This choice is mainly made in order to narrow the scope, but also because of the way that adding a full replica influences fulfilling these requirements in different ways. The availability aspect, as well as organisational aspects and cost, are just as important and can be included at a later stage using the same technique as for bit integrity and confidentiality. The organisational aspects could also use the criteria from the Trustworthy Repositories Audit & Certification (TRAC)[6] for disposition of requirements and relevant BR characteristics.

Looking closer at integrity, we find that authenticity is not relevant in connection with a BR which only is concerned with bits, and rendering and transformation is also out of scope. Neither the ISO 27000 series nor TRAC is specific in expressing integrity in terms of bit preservation, although DS/ISO/IEC 27005 annex C has a useful list with examples of typical threats. These are partly included in our list of requirements. However, the risks prevention based on ensuring bit preservation (number of copies, integrity check frequency and independence between copies as described in [4]) needs to be taken into account as well. This gives us the requirements tree as illustrated in Figure 4. It is drawn using the open source mind map tool Freemind.

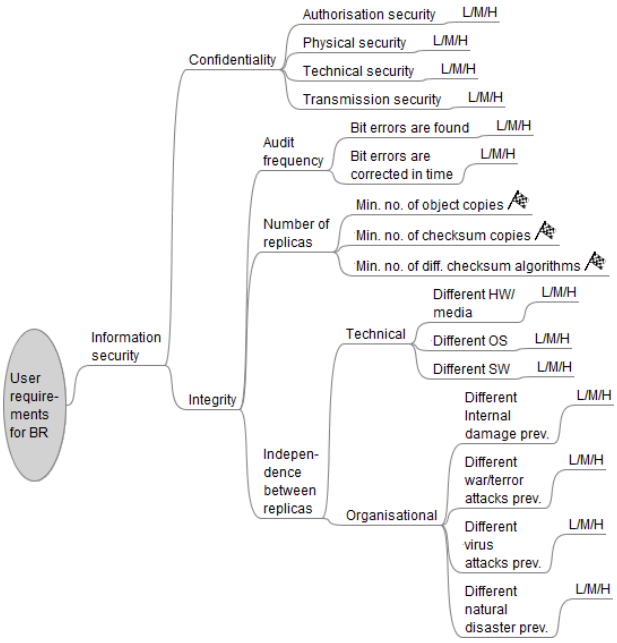


Figure 4. Requirements for a BR (in a mind map).

The branches indicated by a flag symbol are only indirectly included here, in the sense that they are specified as part of the SLA cases which we will define later. The rest of the branches represent importance of requirements which can be measured using an ordinal scale Low/Medium/High. In the following the requirements from the different branches in Figure 4 are explained. For later reference, each requirement is prefixed with an abbreviation number.

According to the ISO standard the *confidentiality* related requirements should be specified to how data is classification in terms of value, legal requirements, sensitivity and criticality. This leads to requirements of preventing the following risks.

C1: Authorisation security violation, which concerns authorisation in all parts of the BR.

C2: Technical security violation which includes e.g. spying via technical means

C3: Physical security violation which concerns e.g. physical access and theft.

C4: Transmission security violation which particularly looks at transmission issues

The *audit frequency* to ensure *integrity* addresses frequency and timely data restoration. This leads to requirements of preventing the following risks:

A1: Bit errors are found which depends on algorithms for detecting errors and timely appliance.

A2: Bit errors are corrected in time which depends on e.g. when corrective actions take place, and how often audit checks are performed held up against mean-time-to-failure for the individual replicas.

The *independence* between replicas is to ensure that *integrity* is not compromised due to similar errors which can corrupt the data in similar ways.

[6] See http://www.dcc.ac.uk/tools/trustworthy-repositories/

Risks to be prevented by differences on the *technical* level are:

IT1: Different hardware/media which concerns both the type of media and vendors of hardware.

IT2: Different operating system which concerns the origin of the operating system, the type, and the vendor.

IT3: Different software ensures that the same error will not occur for several copies due to same error in the software installed, e.g. language interpreter or software for BR application.

Risks to be prevented by differences on the *organisation* level are:

IO1: Different internal damage preventions which concerns internal damage e.g. caused by an operator. For simplicity we have also included errors caused by faults in power supply under this category.

IO2: Different war/terror attacks preventions which e.g. relates to the geographical location.

IO3: Different virus, worms attacks preventions which related to how such attacks are prevented.

IO4: Different natural disaster preventions, where natural disaster can be anything from flood to volcanic activity. For simplicity we have also included errors caused by magnetism or radiation here.

4. EXPERIMENT CASES

To make the final cases for evaluation of bit preservation strategies, we need to define cases for; firstly, the digital material to which we want to make a bit preservation strategy along with the levels of risk prevention that we require. Secondly, a case of a BR implementation which offers different bit preservation solutions along with cases of SLAs defining how the services can be used for the digital material.

4.1. Material Cases

The material cases cover different data material that require different confidentiality and bit integrity levels. In Figure 2 this is the 'material requirements' which are expressed as importance of preventing the risks expressed in the requirements tree (see Figure 4). Each material case is prefixed with an abbreviation number, which will be used as reference in later tables.

M1: Digital born diaries which are highly confidential, and irreproducible.

M2: Digital born images which are open to the public and irreproducible

M3: Digitised books that are open to the public, and reproducible through re-digitisation.

Table 1 shows the requirement which we have estimated for the different material cases. The importance of preventing the risks is L=Low, M=Medium or H=High.

Requirement	Material case M1	M2	M3
Confidentiality			
C1 (author.)	H	L	L
C2 (phys.)	H	L	L
C3 (tech.)	H	L	L
C4 (trans.)	H	L	L
Integrity			
A1 (found)	H	H	M
A2 (corrected)	H	H	H
IT1 (HW)	H	H	L
IT2 (OS)	H	H	M
IT3 (SW)	H	H	M
IO1 (internal)	H	H	M
IO2 (war)	H	H	L
IO3 (virus)	H	H	H
IO4 (disaster)	H	H	M

Table 1. Requirements for digital material cases.

The Table 1 shows that for M3 (digitised material) it is of medium importance to find single errors, but of high importance to have errors corrected, if large volumes and thus investment of the original digitisation are in danger. Loss of data in a war or terror attack is however only viewed as of low importance.

4.2. BR Case

As a case of a 'specific BR implementation' (see Figures 1 and 2), we have selected different pillar implementations and defined characteristics and functions for calculation of requirements probabilities.

4.2.1. Selected Pillars

As basis for a concrete BR we have made examples of pillars used for Danish BR implementation, supplemented with a cloud pillar (e.g. DuraCloud[7]) and a pillar under different law. The pillars are listed in Table 2.

Pillar	Short description
DiCph	Distributed disk system with RAID in org. A in Copenhagen
DvCph	Off-line DVD in org. C in Copenhagen
TpAar	Tape station in org. B in Aarhus (app. 100 km from Copenhagen)
DiAar	Server optimized for robustness in organization B in Aarhus
Cloud	Cloud in unknown organisation
DiAus	Disk based system in org. in Austria

Table 2. Pillars in BR case.

The cloud pillar is interesting because clouds are emerging, and it would be relevant to see what impact a full replica in a cloud could have on bit integrity and confidentiality. A parameter for bit integrity is also the

[7] See http://duraspace.org/duracloud.php

geographical placement, to determine distances between pillars and danger zones pillars are located in. Since Denmark is small which, we have chosen to add a pillar placed in another country. This choice can also affect confidentiality, because of legal issues.

A pillar has many characteristics and changing just one characteristic can mean a different outcome. The naming of the pillars should therefore only be taken as a short abbreviation for some of its characteristics.

4.2.2. Selected Characteristics

The system and pillar characteristics are many. Even in the prototype BR-ReMS the number is about 100. Therefore we will here only explain what they cover generally, illustrated with a few examples, and referencing where further relevant input can be found.

The characteristics included for this case study are partly based on details of the ISO 20005 Annex C on typical threats. More detailed characteristics could be made by adding relevant parts from the ISO 20005 Annex D on vulnerabilities and methods for vulnerability assessment. Note that Annex D is a specialisation of Annex C, or rather Annex C lists the threats that can cause the vulnerabilities.

The ISO standard takes another approach than the one described here, since its aim is not calculations. For calculations, we need parameters from the technical and the organisational perspective, as well as defining them in terms of facts of the implementation. For instance, concerning risk of flood, we need characteristics on if it is in a flood zone, and in this case what organisational and physical prevention procedures that exist.

Additionally, there are characteristics that are specific to active bit preservation (e.g. bit audit frequency, type of checksum algorithm) and the facts on technical details (e.g. on capacity, Mean-Time-To-Failure, expected hardware life time, media technology) and organisational data (e.g. physical location).

4.2.3. Selected Requirements Calculations

Because of the large number of characteristics and the complex interrelations, the calculations are made at varied levels of detail. For instance the IO1 (internal damage prevention) depends on 25 characteristics.

4.3. SLA Cases

The SLA cases represent cases of 'specific use of BR' (see Figure 2) and constitute the alternative solutions for bit preservation. These are therefore the alternatives to be specified and evaluated in Plato.

The SLA cases consist of a pillar combination for the replicas, as well as the type of replica (C=checksum, F=full) that is stored on the individual pillars. Table 3 lists the following SLA cases with choice of pillar combinations and replica types:

S1: As present in DK (except a checksum replica).
S2: Influence of exchange with checksum replica.
S3: Optimised confidentiality in organisation A.
S4: Influence on confidentiality with Cloud replica.
S5: Optimised bit integrity with two full replicas.
S6: Influence of an extra checksum.

	SLA case					
Pillar	S1	S2	S3	S4	S5	S6
DiCph	F	F	F	F	F	F
DiAar	F	C	C	C	C	C
TpAar	F	F				
DvCph			F			
Cloud				F		C
DiAus					F	F

Table 3. Service Level Agreement cases.

For the sake of simplicity we here leave out SLA details on e.g. frequency of bit audits, and we only use one type of checksum e.g. MD5.

5. RESULTS

We will now look at the results we can get from use of the methodology on the simplified case studies. We will firstly look at the results of the BR-ReMS prototype, before proceeding to the actual evaluation using Plato.

5.1. Prototype BR-ReMS Results

The BR-ReMS prototype found that the different requirements were met to L=Low, M=Medium or H=High degree for the different SLA cases. The results are listed in Table 4.

	SLA case					
Requirement	S1	S2	S3	S4	S5	S6
Confidentiality						
C1 (author.)	M	M	H	L	L	L
C2 (phys.)	M	M	H	L	M	M
C3 (tech.)	M	M	H	L	M	M
C4 (trans.)	M	M	H	L	M	M
Integrity						
A1 (found)	M	M	M	M	M	H
A2 (correctd)	M	L	L	L	M	M
IT1 (HW)	M	M	H	L	M	M
IT2 (OS)	H	H	H	L	H	H
IT3 (SW)	M	H	H	L	H	H
IO1 (internal)	M	M	M	L	M	M
IO2 (war)	M	M	L	L	H	H
IO3 (virus)	M	M	H	L	M	M
IO4 (disaster)	M	M	M	L	H	H

Table 4. BR-ReMS results of requirement fulfilment.

Note that the results given here are made independently of specific material cases. It can also be noticed that especially case S4 generally has a very low score on most requirements. The reason is that one full replica was placed in a cloud, where we do not know much about the pillar characteristics. Since the calculations need to account for worst case, we

consequently get the value Low for many of the requirements. Note that if we had more precise knowledge of the cloud pillar characteristics then this picture would probably differ.

The difference between case S1 and S2 was that one full replica was exchanged with a checksum replica. This gives lower score on correction, but also higher score on different software. The reason is that difference in hardware only looks at variations for full replicas, which in this case are placed on the two pillars that differ in software.

The relatively high scores in case S3 are mainly a consequence of having one full replica on highly secured DVDs that are off-line and non-magnetic material. There is also a parameter that the other full replica is handled in house.

It is important to note that these results are only indications. The BR-ReMS is still only a prototype. More granularity and more specific functions are needed to give more precise measures.

5.2. Plato Results

Firstly we make a general evaluation of how well the different six SLA alternatives meet the requirements in general, i.e. not considering specific material cases. The results are given in table 5:

	SLA case					
Rank level	S1	S2	S3	S4	S5	S6
Confident.	1,5	1,5	2,5	0,5	1,3	1,3
Integrity	1,6	1,4	1,5	0,8	1,8	2,0
Total	3,1	2,9	4,0	1,3	3,0	3,3

Table 5. Plato results for SLA cases in general.

The results are found by transforming the BR-ReMS results to a uniform scale between 0 and 5 for each requirement (here using: Low=1, Medium=3, High=5), which Plato uses to give a ranked list of the alternatives. For simplicity only the totals for confidentiality and integrity requirements are included in the table.

The ranking in Table 5 shows that case S3, designed to ensure high confidentiality, has the top score both in total and on the confidentiality level. The case S6, with an extra checksum, is the top score on the integrity level. Finally, the S4 case including a full replica in a cloud is ranked with lowest score, due to the low score in the BR-ReMS.

Now we proceed with the evaluation for the three specific types of digital material. Here we scale the results by comparing the required level of importance with the resulting degree that the requirement is met. The schema for defining scales is given in Table 6. The zero value is based on a decision ***not*** to accept a result where the importance for of a requirement for a specific material case is High, but for a specific SLA case the resulting BR-ReMS probability value is Low.

	BR-ReMS result		
Required value	L	M	H
L	5	5	5
M	3	5	5
H	0	3	5

Table 6. Transforming scheme to Plato scale.

5.2.1. Plato Results for Case M1

Table 7 gives the Plato results for *digital born diaries* (M1). There is only one alternative for the digital born diaries with a utility value greater than zero, leaving case S1 as the optimal solution. The reason that the other cases are eliminated is that there appear zeros for one or more of the requirements, thus the total performance value becomes 0 (stars indicates where such a requirement appeared in the table).

	SLA case					
Rank level	S1	S2	S3	S4	S5	S6
Conf.	2,4	-	-	-*	-*	-*
Integrity	0,6	-*	-*	-*	-	-
Total	3,0	0	0	0	0	0

Table 7. Plato results for SLA cases to M1.

In the case study we actually designed case S4 to fit this material, and case S4 did also have good scores on confidentiality in the general evaluation (see Table 5). Even taking the inaccuracies into account, this is therefore a bit surprising. The detailed reason is that case S4 got Low score for requirement A2 'bit errors are corrected in time' (see Table 4) while the requirement was to have a High score (see Table 1). The same applies for the requirement IO2 'Different war/terror attacks preventions'. The decision not to accept a Low value for a High requirement therefore has the result of eliminating case S4. This is quite reasonable, when we look at digital born material.

The reason for the Low score on requirement A2 is that one full replica is placed on a DVD in the DvCph pillar (see Table 2), which in our example is only properly checksum checked every 2 years. Even though a separate checksum is offered for voting, there is relatively high risk that the full replica on the DvCph pillar may also be damaged, in cases where the full replica on the DiCph pillar is found to be with error. The reason for the Low score on requirement IO2 is that the two full replicas are placed only one kilometre apart.

If we had chosen only to give positive values in the scores (see Table 6), the result would have been different and case S4 would have been chosen. In a real life situation the choice of zero would be reasonable, and the result should therefore instead lead to a new evaluation, where e.g. a full TpAar replica was added to the SLA. Note, that in some cases, only minor changes in a SLA, e.g. frequency of integrity check on a specific pillar, could make a difference for the result.

5.2.2. Plato Results for Case M2

Table 8 gives the Plato results for *digital born images* (M2). The winning alternative for digital born images is case S6. Cases S2, S3, and S4 are eliminated for the same reasons as for the M1 (High requirement value for A2). This leaves the cases S1, S5 and S6.

	SLA case					
Rank level	S1	S2	S3	S4	S5	S6
Conf.	1,0	-	-	-	1,0	1,0
Integrity	2,5	-*	-*	-*	2,9	3,3
Total	3,5	0	0	0	3,9	4,3

Table 8. Plato results for SLA cases to M2.

It is quite reasonable that case S6 wins over case S5, since case S6 contains the same pillars as case S5, but added with an extra checksum. On the other hand it is not obvious why case S6 wins over case S1, since case S1 has three full replicas, while case S6 has only two full replicas and two checksum replicas. The reason is that case S6 is better protected against war and natural disasters by having a full replica abroad (pillar DiAus). Details in the result also show that case S6, because of the extra voter, has a better score than case S1 on requirement A1 'Bit errors are found'. However, because of the inaccuracies in this study, this should *not* lead to a conclusion that an extra checksum is better than having three full replicas.

5.2.3. Plato Results for Case M3

Table 9 gives the Plato results for *digitised books*. All three alternatives S1, S5, S6 are winners as equally good alternatives for digitised books.

	SLA case					
Rank level	S1	S2	S3	S4	S5	S6
Conf.	2,5	-	-	-	2,5	2,5
Integrity	2,2	-*	-*	-*	2,2	2,2
Total	4,7	0	0	0	4,7	4,7

Table 9. Plato results for SLA cases to M3.

Cases S2, S3, and S4 are eliminated since we also here required high score for A2 'bit errors are corrected in time'. It can here be noted that case S3 would win in the case of M3, if the score for A2 had not been zero. All other requirements would then have score 5.

A more highly evolved BR-ReMS, with more granularities and details, would likely produce different results, which could lead to choice of a case. Adding requirements on cost and availability, will also change the similar performance values. The reason is that digitised material available for the public, most probably will have requirements of relatively low costs and fast access to material e.g. via a pillar with distributed architecture with high CPU power per data volume.

6. DISCUSSION

As pointed out several times, it is the methodology that is the result of this article. The results of the case studies only illustrate the use of the methodology, rather than giving real life trustworthy results. In order to get better results, there still is work to be done on the requirements aspects such as costs, detail and coverage of pillar characteristics, better BR-ReMS functions for calculating fulfilment of requirements, and more extended use of facilities in Plato.

Requirements could be further developed using the ISO 27000 standard, but could also be based on TRAC including organisational trust, or other models. It should be noted that the methodology does not try to be a substitution for audits following such standards. The calculations made in the BR-ReMS can only give approximations, no matter how detailed it gets. It is meant as a support in evaluation of a bit preservation strategy. Audits of whether pillar characteristics hold should be supplements possibly required in a SLA.

Additional refinement, both on requirement level and pillar characteristics, could be made for issues like encryption, compression, checksum checks using different checksum types etc. Note that these could also be added on the requirements level, if for example an organisation has a policy that *no* digital born material may be encrypted. Use of Plato could also be much more advanced for such cases, e.g. weighting the non-encryption requirement high compared to other requirements. Furthermore, granularity of values for requirements and results could be enhanced to give more nuanced analysis.

Refinement of the functions for requirement fulfilment will be a subject for discussion. Firstly, detailed and possibly automated calculations can easily become too complex to audit, and too rigid to handle inclusion of new aspects. Secondly, different approaches to calculate whether bit audits are done as frequently as needed may give a different outcome. The calculation will probably be based on measures like Mean-Time-To-Failure where it can be debateable how much we can trust such measures.

The level of refinement should also take e.g. hardware/media migrations and upgrades of software into account. If the level of details for characteristics and requirements are too high, it will be hard to make e.g. migrations without re-negotiating all SLAs using the pillar in question. The best solution would be, if a migration plan could be based on re-calculations of characteristics to see whether it would have any negative affect on them. In this case the migration could take place without any re-negotiations.

7. CONCLUSION

The presented methodology has been shown to be useful as an aid to evaluation of alternatives for a bit preservation strategy. Even for the simple case study, with little granularity in requirements and results, and with a BR-ReMS prototype with little refinement, we could produce results that pointed out weaknesses in the SLA cases covering different pillars and characteristics.

The planning tool Plato helps in the analysis of the results. Without Plato, it would have been much more difficult to analyse the results of the BR-ReMS.

The BR-ReMS has also proven useful, at least in the way it structures characteristics for a BR. There may be other approaches to define requirements which the BR-ReMS also can support.

Even though the methodology has been shown to work, there is still a lot of work to do on requirement specification including standards like TRAC, ISO, DRAMBORA[8], and work on detailing the BR-ReMS on characteristics and calculations on requirements specification. Furthermore development of more detailed requirements in Plato will enhance the outcome of using the methodology.

Further work will also study how the methodology can assist consumers in choice of bit preservation strategy and formulation of SLAs, as well as how it can assist service providers in long term operation of parts of a bit repository fulfilling SLAs.

8. ACKNOWLEDGEMENTS

Part of this work was supported by the European Union in the 6th Framework Program, IST, through the PLANETS project, contract 033789.

9. REFERENCES

[1] Becker, C., Kulovits H., Rauber A., Hofman H. "Plato: A service oriented decision support system for preservation planning", Proceedings of the Joint Conference on Digital Libraries, Pittsburgh, USA, 2008.

[2] DS/ISO/IEC 27000-27007, first edition, 2009.

[3] Kulovits H., Rauber A., Kugler A., Brantl M., Beinert T., Schoger A. "From TIFF to JPEG 2000? Preservation Planning at the Bavarian State Library Using a Collection of Digitized 16th Century Printings", D-Lib Magazine Vol. 15 No. 11/12, 2009.

[4] Rosenthal, D.S.H. "Bit Preservation: A Solved Problem?" *Proceedings of the* International Conference on Preservation of Digital Objects, London, United Kingdom, 2008.

[5] Rosenthal, D.S.H., Robertson, T., Lipkis, T., Reich, V., Morabito, S. "Requirements for Digital Preservation Systems, A Bottom-Up Approach", *D-Lib Magazine Vol. 11 No. 11*, 2005.

[6] Zierau, E., Kejser, U.B. "Cross Institutional Cooperation on a Shared Bit Repository", *Proceedings of the International Conference on* Digital Libraries, New Delhi, India, 2010.

[8] See http://www.repositoryaudit.eu/

PRESERVATION PLANNING: A COMPARISON BETWEEN TWO IMPLEMENTATIONS

Peter McKinney
National Library of New Zealand
Te Puna Mātauranga o Aotearoa
PO BOX 1467 Wellington, NZ

1. INTRODUCTION

This paper examines preservation planning as it is implemented within the National Library's preservation repository (Rosetta) and compares it directly to the PLATO tool created as part of the PLANETS project.

Preservation planning is both a business precondition and the systematic framework defining any preservation action. At the National Library of New Zealand Te Puna Mātauranga o Aotearoa, preservation planning is embedded within the Rosetta system.

For the Library, the challenge can be stated simply: preserve New Zealand's digital documentary heritage. With no limitations or control over the format of the content that is collected and preserved, The National Library of New Zealand Te Puna Mātauranga o Aotearoa (NLNZ) has 'issues' to resolve before the long-term preservation of digital collections can be assured. Solving these and other problems is the responsibility of the National Digital Heritage Archive and a significant step has been taken through the development of the Rosetta preservation repository system in conjunction with Ex Libris Group.[1]

1.1. Preservation Planning context

Library policies at the highest and most base levels have created an institutional context that imposes itself on the requirements for preservation planning. A few key salient factors are outlined here and will be developed further in the fuller paper.

The National Library can, and does accept all formats. It collects content, not 'perfect' formats. All materials collected through legal deposit are ingested into the preservation repository essentially as is, and the current policy of the NDHA is to not transform content into preferred formats on ingest.

Risk analysis is situational and characterized by understanding institutional capability. The Library does not use sustainability factors for generating a risk view of its content. The range of formats ingested along with the imperfect nature of identification and characterisation tools necessitates the creation of a risk profile based on the Library's ability to render content. A two-tier view allows us to see exactly what content the Library can render through a systematic relationship between formats and applications and can take account of any specific properties of files that impinge on that relationship [3].

The basic policy of the National Library is that we will only be taking action on content if we are truly required to. While this may seem to be a self-evident statement, it is important to understand that by basing risk on institutional capability the Library is not beginning preservation planning on 'what if', but rather is working on a 'we have to' statement. We believe that 'what if' is linked to the use of risk analysis that is based on projections using 'sustainability criteria' [1], rather than definite capability tracking.

1.2. Preservation planning

The ultimate goal of preservation planning for NLNZ is for the plan to become a defined course of action. That is, it becomes the unchangeable template of action against which every file that matches its criteria follows. To get to that point, a number of critical elements need to be in place and a structured workflow must be successfully negotiated.

It is these elements and stages that will offer the initial comparison with the PLATO functionality. Preservation planning in Rosetta was not created within a vacuum. A great deal of time was spent modelling both the Library's and Ex Libris's expectations and testing these against the detailed flow developed by the PLANETS project and are embodied in the PLATO tool [2].

This paper will deliver not only a line-by-line comparison of the elements both types of planning have identified as required, but will also explore the institutional background behind the major points in both planning frameworks, particularly at points of difference.

The paper will undertake a comparison in the following areas:

- The place of the frameworks within the lifecycle of digital objects
- Workflows
- Evaluation of plans
- Presentation of plans to decision-makers

It is clear that there are a great number of similarities between PLATO and the Rosetta planning framework. Both are grounded by a focus on the presentation of solid information to decision makers from which the best path forward can be decided upon. However, initial work has identified some divergence, characterized mostly by differences in emphasis and the timing of some of the stages.

For example, by virtue of being an active preservation repository dealing with heritage items, the place of planning within Rosetta appears to be more tied to mitigating an occurred risk. This has ramifications on lifecycles of plans and the environment they are created within. The paper will explore whether this conjecture is valid.

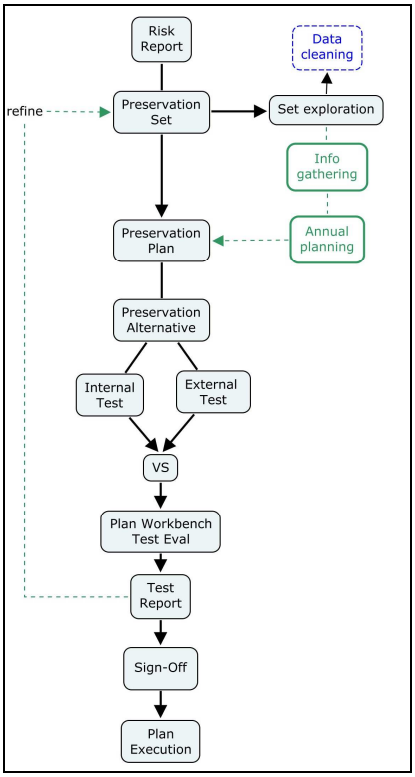


Figure 1. Overview of preservation planning in Rosetta

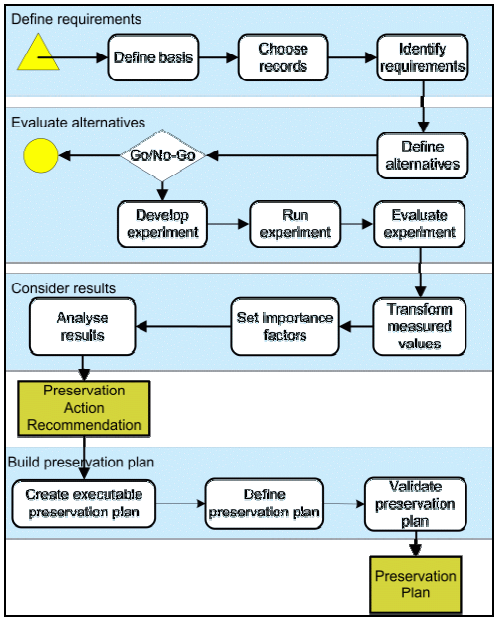


Figure 2. Overview of PLANETS Preservation Planning Workflow[2]

2. INITIAL REFERENCES

[1] Arms, C. & Fleischhauer, C. "Digital Formats: Factors for Sustainability, Functionality, and Quality", *IS&T Archiving 2005 Conference*, Washington, D.C.

[2] Becker, B., Hannes, K.,Guttenbrunner, M., Strodl, S., Rauber, A. & Hofman, H. "Systematic planning for digital preservation: evaluating potential strategies and building preservation plans", *International Journal on Digital Libraries, December 2009*

[3] De Vorsey, K. & McKinney, P. "One man's obsoleteness is another man's innovation. A risk analysis methodology for digital collections", *IS&T Archiving 2009*, Virginia, 2009

[4] Kulovits, H., Rauber, A., Kugler, A., Brantl, M., Beinert, T. & Schoger, A. "From TIFF to JPEG 2000? Preservation planning at the Bavarian Stea Library using a collection of digitized 16th century printings", *D-Lib Magazine*, 15:11/12, 2009. http://dlib.org/dlib/november09/kulovits/11kulovits.html

[2] This diagram is taken from the PLATO website. http://olymp.ifs.tuwien.ac.at:8080/plato/help/workflow.html.

Session 5b: Processes and Best Practice

QUALITY INSURANCE THROUGH BUSINESS PROCESS MANAGEMENT IN A FRENCH ARCHIVE

Marion Massol **Olivier Rouchon**

CINES
Long term preservation department

ABSTRACT

This paper outlines the recent initiative run at CINES, a national IT datacentre for French academic researchers, to formalize the business processes of its department dedicated to the long-term preservation of digital objects, which is at present one of the very few operational long-term preservation platforms in France for the public sector and Higher Education and Research in particular.

One of the strategic goals of this organization is the evaluation and assessment of service quality. The processes formalization activity – coupled with an external audit and an ITIL approach – highlighted the department good practices, gaps and weaknesses.

The processes global map and most of the twenty detailed process maps have been put together to support the team in its documentation goal and are available online on the institutional web site, along with the CINES specific rules for archival processes formalization based on standards such as ISO 9001 or ISO 14721.

This experiment has revealed that such a process approach can to be an excellent mean to structure and plan for an efficient implementation of the preservation strategy as well as an opportunity to improve service quality, which is actually the final objective of the digital long-term repository of the centre since it's aiming at the future ISO 16363 certification.

1. BACKGROUND

CINES (Centre Informatique National de l'Enseignement Supérieur, a national datacentre for the higher education and research community) is a public French organization known worldwide for its HPC (high performance computing) activities.

CINES was also entrusted with a long-term preservation of electronic data assignment. Three types of digital documents are secured on the archiving repository called PAC (Plateforme d'Archivage du CINES) for the years to come:

- Scientific data generated from observations, measurements or computation;
- Heritage data like PhD theses, educational data or pedagogics, publications or scientific digitized books;
- Administrative data from French universities: personal records…

A department dedicated to electronic archiving was created in 2003 and PAC was one of the first long-term preservation platforms deployed in production for the public sector in France. At present, more than ten FTEs work in the long-term preservation department, who have developed their own information system and participate in French and European projects.

In 2008, a quality initiative began with the objective of getting a full, transverse certification of the service covering archiving, technical, organization and service aspects. To date, CINES has been involved in the Data Seal of Approval (DSA) accreditation provided by the DANS (Data Archiving and Networked Services), a Dutch organization, and is still a member of the DSA Editorial Board. The long-term preservation department has also been audited by senior external consultants, in order to identify its strengths and weaknesses, and to prepare the certification within an accepted time frame. The strategy adopted by the Management to achieve its quality target relies on risk management (based on the DRAMBORA method) and a solid business processes formalization.

2. THE PROCESS APPROACH: WHAT IS IT ABOUT?

To increase readability and understanding, business processes formalization is based on graphical representations, using standard shapes and connectors to describe a sequence of events, alternatives or activities. Data sheets complement the graphics and describe all formalized objects, to comply with the international standards ISO 9001 and 9004.

For some processes, the methodology used for the formalization is a bottom-up one; for the others, it is a reengineering approach. Elaboration and validation of business processes systems have been done by iterations: experts are interviewed, process system managers represent processes, then, the whole archiving team and the management approve process maps before their publication on line.

The Business Process Analysis (BPA) approach and the work presented in this paper have been validated by French senior consultants specialized in long-term preservation issues and New Information and Communications Technologies (NICT).

3. EXISTING STANDARDS: WHY ARE THEY INSUFFICIENT?

The international standard ISO 9001 requires putting together specific documentation for each identified business process. All the characteristics of the process, a description of the workflow and identification of people who are responsible for its execution have to be summarized in a data sheet. But, CINES has gone beyond these requirements, as it didn't stick to a limited list of classified processes. Indeed, a set of process maps has been documented: a general map allows a global view, while a set of more detailed maps ("macro" and "detailed" maps) focuses on a particular formalized object. So, this type of formalization gives CINES a real basic reading of its network process.

The international ISO 14 721 standard is the main archival reference used by business analysts to formalize their archival processes. The OAIS model is very interesting because it proposes a functional view of the Archive. Six functional entities (Ingest, archival storage, access, data management, preservation planning and administration) include about thirty functions. However, it is not always easy to understand how powerful this standard is, and the complex interactions between the different functions described. The vision proposed by OAIS is vertical. On another hand, the business process approach is based on the split of the OAIS functions into objects called "activities" and their sequencing through relationships between them. The diagram shown on the opposite highlights one of these semantic links for the access process.

Furthermore, the OAIS standard focuses mainly on archiving activities. The business processes approach of CINES integrates a few more aspects not yet described in the ISO 14 721 but which are vested in the Archive. For example, as with all French public organizations, CINES has to follow special procedures for procurement, and must report on its mandates to its supervisory and funding body. These two processes – and a few others, such as risk management, metrics and results appraisal, documentation management or prospective aspects – have been documented along with the functions described in OAIS.

Then, business process approach integrates few aspects from other standards. For example, security issues have been dealt through the ISO 27 001 requirements. The SEDA, a French exchange standard which formalizes relationships between the Archive, producers and users, has been used to improve and extend the OAIS model, particularly around the archive ingest and access issues.

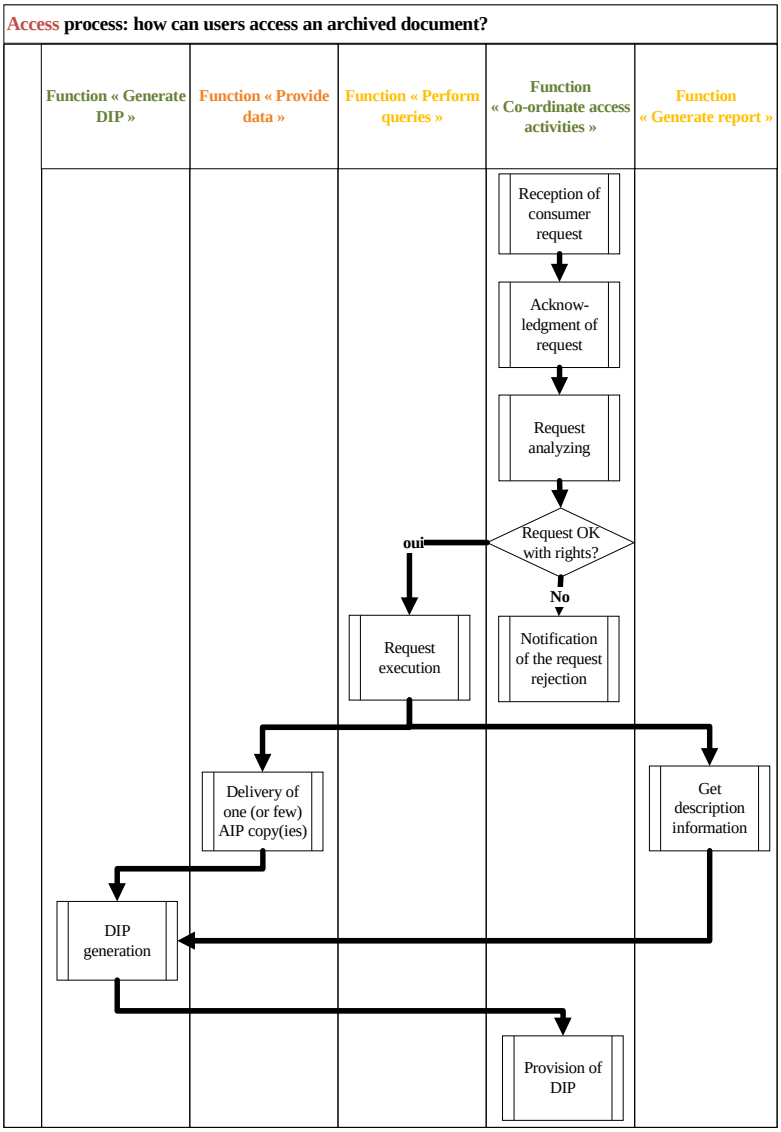


Figure 1. Access business process: creation of relations between activities as in OAIS model.

4. ADVANTAGES OF THIS INITIATIVE: WHY DOING IT?

The business process approach is interesting because archiving issues are handled and entirely integrated into a greater scope. Thanks to the integration of ISO 9001 and other standards in the OAIS model, the business approach is complementary to classic initiatives.

So, this different focus allows a permanent auto-evaluation of the Archive with as many different points of view as different standards in the business process referential. The ability to realize its own evaluation is one of the prerequisites that an Archive has to meet to be certificated with, for example, the upcoming ISO 16363 standard for audit and certification done by CCSDS and ISO committees.

The business process documentation is the spine of the Archive as it answers inevitable questions around its activity: it designates the owners ("who") of the processes, describes the triggers ("when") or the rationale ("why") for their usage, grades their importance or priority, and shows when there is a competitive advantage. As a consequence, during a certification initiative, such a document is considered by any auditor as essential.

The documents produced can help facilitate comparisons with other similar structures: as an example, the CINES risk management seems to be more focused on the mitigation activities rather than on the actions following the occurrence of the risk, which is a different approach from other French public institutions.

Economical benchmarking, various comparisons and better comprehension of your own internal running, as well as process repeatability are the building blocks of a more structured deployment strategy. Global policy comes in a variety of business process targets. This approach doesn't fit with a hierarchical partitioning of function groups or teams, as engineers, archivists and other people involved have to work together to execute their shared processes as smoothly as possible. The business process approach gives elements to understand interactions between different parts of the organization and to facilitate interdepartmental cooperation.

Furthermore, business processes documentation proved to be an important vector of knowledge management and dissemination by improving internal communication, as it helps to bring newcomers up to speed on the Archive operation. Preserving digital objects is a key objective of repositories, which can only be achieved in the long term by preserving the high level of expertise that the team responsible of the Archive has acquired. When budgets shrink and turnover rate increases in institutions, this should be kept in mind and managing business process documentation should be considered carefully.

This initiative also improved the overall performance of the archival information system, as processes have been assessed, questioned and rationalized; metrics and dashboards have also been improved and fine-tuned as part of the same BPI (Business Process Improvement) exercise.

Last, but not least, transparency on its own processes is a good way to create trust relationships with user communities, supervisory and funding bodies, partners, peers (i.e. other OAIS repositories involved in digital preservation), etc. For this reason, the CINES business process system has been made accessible on the Internet website. This allows discussions between peers around best practices, choices and strategies, or comparisons with other institutions which would have a similar initiative of business process documentation.

5. A SPECIFIC METAMODEL: HOW TO CREATE IT?

The business process system must be understandable by anyone: any person working in the Archive, any member of executive committee of CINES, any external auditor, any member of another Archive, any representative of the supervisory or funding bodies, etc. So, the representation of processes must follow a set of rules: such a collection is named a "metamodel". Strictly, a metamodel is the representation of a special point of view on models. A model is an abstraction of phenomena in the real world; a metamodel is yet another abstraction, highlighting properties of the model itself. A model conforms to its metamodel in the way that a computer program conforms to the grammar of the programming language in which it is written.

Any formalization is the expression of a particular point of view of a system, so a metamodel has to be created to reflect it. As CINES did not have any BPA tool with a default metamodel, one had to be developed.

UML is a standardized general-purpose modelling language used in the whole world for information systems specification. But it restricts the scope of formalization as focuses on information system, whereas business processes incorporates strategic aspects and policy.

A methodology for the formalization of business processes exists: BPMN (Business Process Management Notation). It's the only future standard existing but it is not mature enough or adopted yet. A criterion for choosing a good description language is, for CINES, its intuitive aspect. Any reader should be able to understand the signification of process maps without having to consult specific documentation. The BMPN is not used by a large community because of the lack of transparency and intuition. Furthermore, although it's a very strict notation, it can provide multiple frameworks to solve a specific problem. For CINES, this last aspect appeared to be a show stopper.

It is very difficult to find metamodels which develop clear sets of rules about representation and interactions

between objects. Apart from BPMN, there are no standards on that formalization issue or real sharing of metamodels: even BPA software vendors don't provide their default metamodel.

The metamodel adopted for the formalization of archiving processes in the CINES is based on the identification of three object-types: processes, sub-processes and activities. A process is, according to the ISO 9001, an object with a semantic meaning that transforms an input into an output element. This process consists of a set of sub-processes. And the latters are sequences of activities. This choice of representation is simple and flexible enough to be juxtaposed with others for a comparison exercise.

The detailed rules description can be found online as the metamodel is documented on the CINES website.

6. THE CINES PROCESS SYSTEM: WHAT DOES IT LOOK LIKE?

CINES has identified twenty processes in its archival system. By the time of writing this text, seventy percent of them have been documented. For the iPRES2010 presentation, CINES will have spent a year and a half working on this formalization initiative, and we might have completed the whole business process system.

The diagram 2 shown alongside simplifies the detailed map of one of the fourteen processes which have yet been documented, the access process.

In accordance with the ISO 9001 requirements, CINES has defined the scope and breakdown of each process taking into account the following criteria:

- transversality, exceeding the boundaries of a function or activity;
- simplicity, manageable number of interactions with other processes and activities correlated in the process;
- completeness of the network process;
- consistency of interactions;
- clarity for stakeholders internal and external process.

Furthermore, CINES complied with the ISO 9001 requirements for process classification. The French national organization for standardization AFNOR, who also represents ISO at a national standpoint, published the documentation FD X50-176 in 2005.

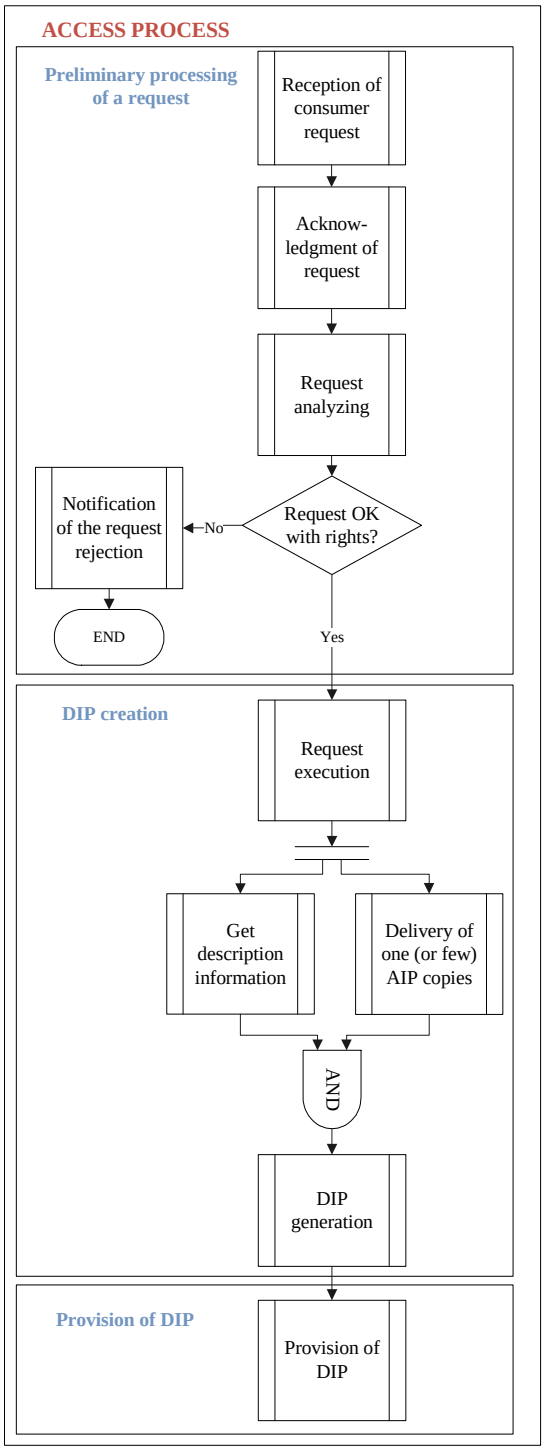


Figure 2. Simplified access process map.

This reference provides a list of process management tools and, in accordance with the ISO 9001:2000, proposes a classification of processes in three groups:

- An implementation group of processes (showed in blue in the global map): based on the ISO 14 721 functionalities, these processes describe the core business;

- A second group with processes that support implementation processes (green processes in the global map below): they gives the means to achieve the long preservation activity and evaluate results. Unlike some business analysts, CINES has chosen not to distinguish these two sub-families;
- And a third process group that manages the whole system of processes (they are formalized in red on the global map).

Such a classification is recursively valid, as sub-processes and activities can be split the same way.

The diagram 3 shown on the next page describes the global processes map.

7. THE DAILY IMPLEMENTATION OF THE PROCESS SYSTEM: HOW DO WE USE IT? HOW IS IT UPDATED?

At present, CINES is still formalizing the business process system. Once a first version of the global processes map established and validated, each process is assigned to and managed by a member of the Archive. The person responsible for a process is member of the internal experts committee. As such, he has been interviewed for the initial formalization of the process and during its updating. This strategy promotes implication and strengthens personal expertise of each team member.

Interconnections between the documented objects imply a regular updating of all the processes already formalized. In this step of the process system specification, process pilots are essential: they ensure the consistency of their own part, while the process system manager is responsible for the total adherence.

This work is a means to encourage experts to consider their work methodology: this action takes the opposing course to routine. Furthermore, with the detection of areas of improvement, this initiative provides assistance for the implementation of global solutions.

8. CONCLUSION: WHAT ARE THE LIMITS OF THIS FORMALIZATION?

In process formalization, two types of projects are identified: mapping and modelling ones. Mapping projects are generally intended only to specify the Archive operation. Unlike modelling ones, these projects don't ensure uniqueness of objects. Adding a description allows a start in a quality initiative such as the ISO 9001 one. With the security on uniqueness of represented objects, modelling projects go further and add in-depth analysis and automation of workflows.

CINES has a specific position because it has neither its own BPA tool (for representation of processes) nor any workflow engine (to execute processes). Consequently, the work presented in this paper is not strictly a modelling project. Nevertheless, maps (done with a specific map tool) and a relational database ensure, with manual controls, uniqueness of objects and consistency of the process system. To avoid any confusion and distinguish a possible switch to real BPA, CINES talks about "maps" rather than "models".

Business process formalization is a step in the quality initiative of CINES. This work is still in progress, but publications are regularly available on the CINES website. As a result of a first external audit of the Archive, the specification of business processes, as well as other elements of the action plan which was put together to work out the gaps identified, has been deployed. As soon as the stabilized referential for audit and certification is chosen, the Archive will prepared itself to this next step of quality process.

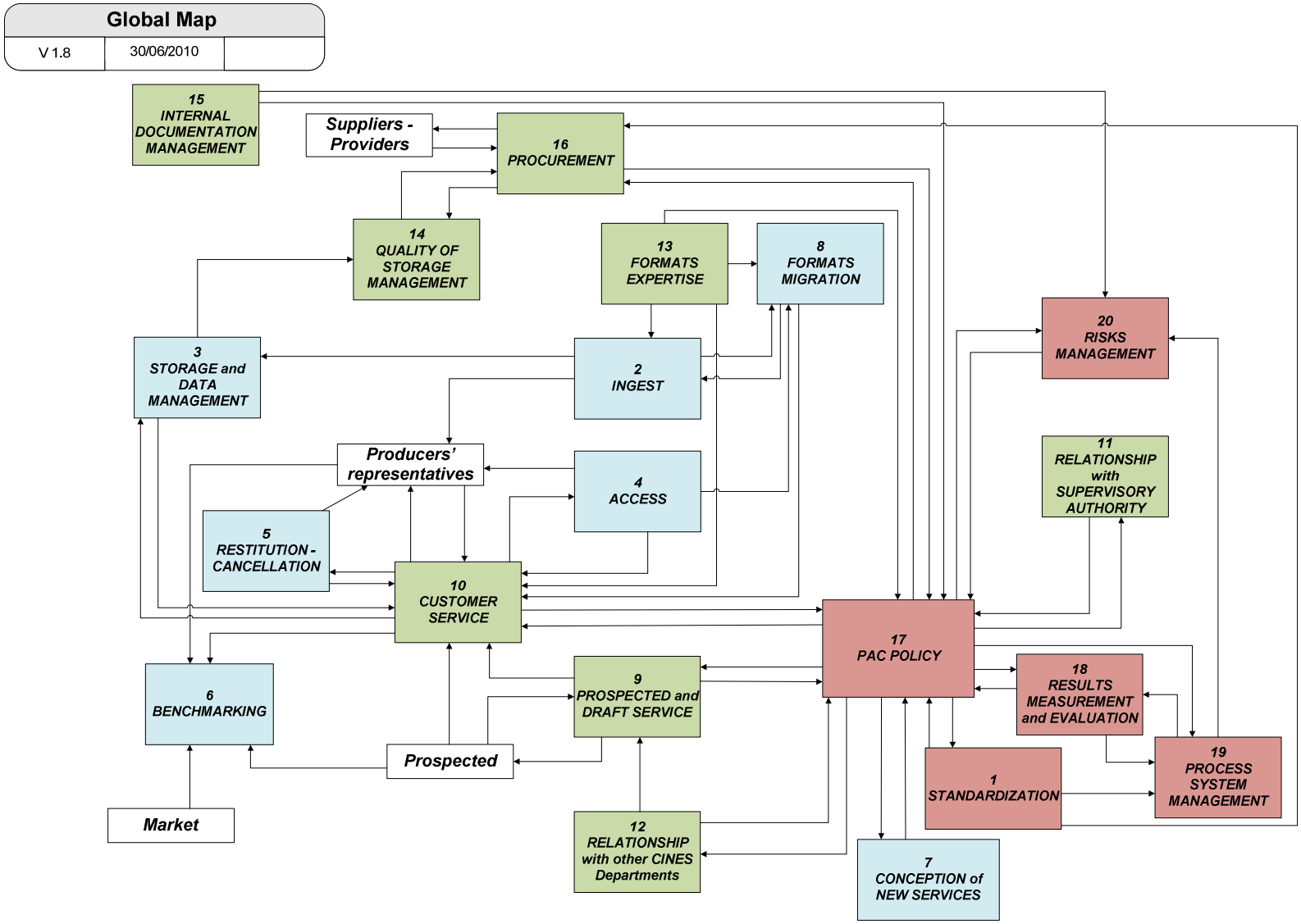


Figure 3. Business process system of CINES.

9. REFERENCES

[1] AFNOR X578 :2005, E. Segot, *Fascicule de documentation FD X50-176, outils de management : management par les processus*, Editorial Afnor, France, 2009.

[2] AFNOR Z40F, M. Cathaly, *NF Z42-013*, Paris (France), Editorial Afnor, 2009.

[3] Calderan, L., Hidoine, B., Millet, J. «*Séminaire INRIA (2*-6 *octobre 2006)* », *Pérenniser le document numérique*

[4] CCSDS 650.0-B-1, *ISO 14 721:2003, Reference Model for an Open Archival Information System (OAIS),* blue book, 2002.

[5] CCSDS 652.0-R-1, *ISO 16 363: Audit and Certification of Trustworthy Digital Repositories,* draft recommended practice (red book) issue 1, 2009

[6] CINES website:
http://www.cines.fr
http://www.cines.fr/spip.php?rubrique4

[7] Data Seal of Approval:
:http://www.datasealofapproval.org/

[8] Direction générale de modernisation de l'Etat (DGME), *Référentiel général d'interopérabilité, version 1.0*, 2009.

[9] DRAMBORA (Digital Repository Audit Method Based On Risk Assesment): http://www.repositoryaudit.eu/

[10] Garminella, K., Lees, M., Williams, Bruce. *Les bases du BPM pour les nuls*, Editorial Solftware AG, Hoboken (USA), 2008.

[11] *ISO 9001:2000 (X50-131),* Editorial Afnor, France, 2000.

[12] *ISO 9001:2005,* Editorial Afnor, France, 2005.

[13] Ministère délégué au budget et à la réforme de l'Etat (direction Générale de la modernisation de l'Etat), Ministère de la culture et de la communication (direction des Archives de France), *Standard d'échange de données pour l'archivage : transfert – communication – élimination – restitution*, Paris, 2010.

[14] Ministère de la Défense, *P2A – politique et pratiques d'archivage (sphère publique)*, Paris, 2006.

[15] Ourouk consultants:
http://www.ourouk.fr

[16] RLG. *Trustworthy Repositories Audit and Certification: Criteria and Checklist (TRAC), version 1.0*, Chicago, 2007.

ARCHIVING ARCHAEOLOGY: INTRODUCING THE GUIDES TO GOOD PRACTICE

Jenny Mitcham **Kieron Niven** **Julian Richards**

Archaeology Data Service

ABSTRACT

This paper looks at some of the domain specific preservation challenges faced by the Archaeology Data Service and how we work with these in order to maximise the re-use potential of the data that we archive. It looks in particular at one of the mandatory responsibilities of an Open Archival Information System (OAIS) and how we try to ensure that the data that we present to our designated community is 'independently understandable'. The paper introduces the collaborative 'Guides to Good Practice' project which aims to provide data producers with the guidance that they need in order to create data that is well documented and thus suitable for archiving and re-use. This Mellon Foundation funded project carried out in association with Digital Antiquity in the United States is now in its final stages and includes comprehensive and practical advice for data creators plus a number of case studies which demonstrate the real practical application of the Guides.

1. INTRODUCTION

The Archaeology Data Service (ADS) was founded in 1996 for the purpose of preserving digital data produced by archaeologists based in the UK, and making it available for scholarly re-use. The ADS was initially established as part of the Arts and Humanities Data Service (AHDS), with sister services covering other disciplines within the arts and humanities. Data are archived to ensure long term preservation, but they are also made available free of charge for download or via online interfaces to encourage re-use.

2. ADS AND OAIS

The digital archive at the Archaeology Data Service was established several years prior to the acceptance of the Open Archival Information System (OAIS) model as an ISO standard. ADS archival procedures and policies have evolved over time as the organisation itself and the wider world of digital archiving has grown and matured. We have now adopted the OAIS model and retrospectively tried to map our archival practices to it, looking in particular at data flows and at the six mandatory responsibilities. This has been an interesting process. Some of the OAIS mandatory responsibilities are easier to comply with than others. The ones which we have found most challenging are (perhaps unsurprisingly) the ones which we have the least control over. In particular where they relate to how the data producers create their data and prepare it prior to archival deposition with us.

OAIS states that an archive should:

> "Ensure that the preserved information is independently understandable to the user community, in the sense that the information can be understood by users without the assistance of the information producer." [1]

This of course is not just up to the archive itself but will inevitably involve some input from the data producer as they are the ones who have the greatest understanding of the data in question and are best placed to provide suitable metadata and other crucial contextual information. Metadata isn't always something which can be generated in retrospect. In many cases it is essential that the metadata is created while project data is being actively generated and processed. It is at this point that creators have the clearest idea of what information each file contains, where it was collected, how it was collected and how it was subsequently processed.

3. THE DISAPPEARING SPIRAL

Take for example a project in 2004 to look for an elusive 'spiral' reportedly carved into the rock on one of the stones of the ancient stone circle at Castlerigg in Cumbria, England. The project team from the Universities of Durham and Bristol used the relatively novel techniques of 3D laser scanning and ground based remote sensing in order to reconstruct the 3D surfaces with millimetre and submillimetre accuracy [2]. These techniques can produce high quality images which can be analysed with a much higher level of objectivity than

more traditional rock art recording methods such as wax rubbings and scale drawings. The team didn't find the spiral, suggesting that perhaps if it ever existed it was painted rather than carved on to the rock. In terms of research, this negative result is just as valid as a positive identification and the resulting point cloud and surface model data was archived with the ADS so that future archaeologists can make use of it in whatever way they like.

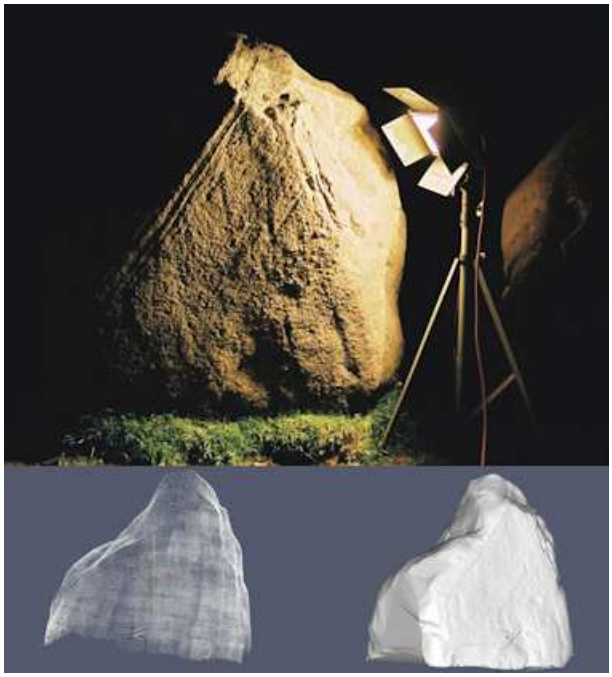


Figure 1. Fieldwork in progress on stone 11 of Castlerigg Stone Circle (top), point cloud and solid model created from the laser scanning data and archived by the ADS (bottom) © University of Durham

Perhaps a researcher some years down the line will want to return to the Castlerigg data files and continue the search for the 'spiral'. In order to fully assess the data from the original fieldwork and reprocess it they would need to know exactly how the 2004 fieldwork was carried out: what equipment was used, the point density on the object, which processing routines were carried out and what software was used. Even information about the date and time of the scan and the weather and light source could be useful. This is the sort of information we should be receiving as part of a Submission Information Package (SIP) so that we can ensure the data has enough contextual information alongside it to make it both understandable and useful. But how can we ensure that we always get what we require?

4. QUESTIONS AND CHALLENGES

Another and perhaps one of the biggest domain-specific challenges that we face as an archive for archaeological data is the range of file types that we are asked to ingest into our archive. A number of the projects we archive (such as that described above) feature cutting edge research using new and innovative technologies. As well as standard file formats that can be found in the majority of archives (documents, images, spreadsheets), we also have to deal with a diverse range of project outputs (maritime and terrestrial geophysics, geographic information systems (GIS), photogrammetry, lidar, virtual reality and more). The resulting files are often large in size and can come in a huge variety of proprietary and binary data formats. Finding ways of preserving these sorts of data can be a challenge. How do we get people to submit data in formats suitable for preservation? Which file types are we able to deal with and what levels of metadata need to be supplied in order to make the data 'independently understandable' to our designated community and thus suitable for re-use?

5. BACKGROUND TO THE PROJECT

These are questions we have been trying to address over the past few years, through projects such as the English Heritage funded 'Big Data' project[1] and the European funded VENUS (Virtual ExploratioN of Underwater Sites) project [2] and also through our previous 'Guides to Good Practice'[3] publications aimed at data producers.

Figure 2. One of the original ADS Guides to Good Practice, *Archiving Aerial Photography and Remote Sensing Data* (both on-line and hard copy versions)

These Guides to Good Practice were published by the ADS from 1998 to 2002 and were available in hard copy and also free of charge as static on-line publications. They focused on subjects such as excavation,

[1] http://ads.ahds.ac.uk/project/bigdata/
[2] http://ads.ahds.ac.uk/project/venus/
[3] http://ads.ahds.ac.uk/project/goodguides/g2gp.html

geophysical datasets, GIS, Computer Aided Design (CAD) and virtual reality, providing practical advice on the creation, preservation and re-use of digital resources and all including useful sections on metadata creation. They had been well received by the archaeological community at the time, but were in need of an update in order to keep up with the latest methods, techniques and technologies in use in these fast moving fields.

6. THE GUIDES TO GOOD PRACTICE

Building on these existing 'Guides to Good Practice' we have, over the last two years, been working with archaeologists in the US to refresh and enhance this resource. The current project is predominantly being carried out in support of the Digital Antiquity initiative, a Mellon Foundation funded US-based project with teams working at the University of Arkansas and Arizona State University.

Through this new, collaborative project we are in the process of updating and restructuring the original Guides, making them available in an on-line wiki environment[4] to allow easy and quick collaboration and also more frequent future updates. In order to keep pace with the wide range of techniques that archaeologists use, we are also including new subject areas such as 3D laser scanning, lidar and photogrammetry (Table 1).

Updated Guides	New Guides
Aerial Survey	Marine Remote Sensing
Geophysics	Laser Scanning
Geographic Information Systems (GIS)	Photogrammetry
Computer Aided Design (CAD)	Satellite Positioning Systems
Virtual Reality	Polynomial Textual Mapping (PTM)

Table 1. The updated and new data types and technologies covered in the new Guides to Good Practice series

Basic Components
Documents and Texts
Databases and Spreadsheets
Raster Images
Vector Images
Digital Video
Digital Audio

Table 2. The 'Basic Components' covered by the new Guides

As well as these technology-specific guides, we have also concentrated on a set of 'Basic Components'. These are the common digital objects that often appear in an archive that is deposited with us, regardless of the nature of the project or the technologies used – primarily textual reports, digital photographs, databases and spreadsheets and occasionally digital audio or video files (Table 2). As these basic components are ones which the majority of data producers will need some guidance on, they have been separated out and are linked to from appropriate places in the other Guides.

In order to create these guides we have invited the original authors (all specialists in their particular fields) to review and update the content. New authors from both the UK and US have also been drafted in to contribute. Once the Guides have been updated, they will undergo wider review by a panel of experts.

The wiki format of these new guides has a number of obvious benefits. Several authors may work on the material simultaneously with the results being made immediately available to all. The wiki allows for page-level privilege control – so authors will have the ability to edit only those sections that they have permissions to author. For each wiki page it is possible for the editor to view the 'page info' in order to see all the edits that have been carried out. This allows them to keep track of all changes that have been made and view all previous versions.

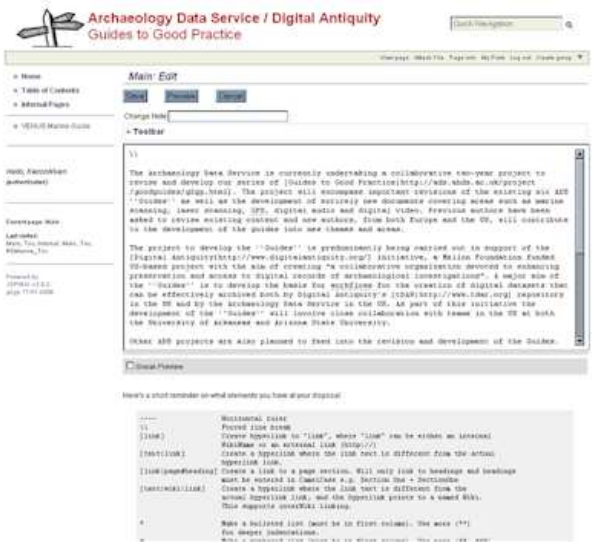


Figure 3. Editing page content in the wiki environment

The Guides will provide data producers with a comprehensive and peer-reviewed set of guidelines explaining how to create data that is suitable for long-term preservation and how to package it up with the correct metadata to ensure it is 'independently understandable'. Different chapters of the Guides target different technologies or groups of files, so users will be able to quickly and easily find the section that is most relevant and useful to them. The wiki format also allows a high degree of interlinking between relevant sections of the Guides making them into a far more interactive resource than previously possible.

[4]http://guides.archaeologydataservice.ac.uk/

Unlike the original Guides to Good Practice series, this new wiki-based publication is not being produced in hard copy form. In recognition of the fact that some archaeologists may want to take a section of the Guides out into the field with them where they have no internet access, and that other users simply may not want to read large quantities of text from a screen, there will be a pre-prepared PDF of each Guide allowing users to download and print-on-demand.

Although the Guides have clearly been written with archaeologists in mind, they do have wider application. Much of the advice contained within them, for example that relating to significant properties, suitable file formats and metadata, will be also be applicable to practitioners in other disciplines. As the project reaches completion the whole wiki will be open and freely available to all.

7. CASE STUDIES

A key component of the Guides to Good Practice project is the inclusion of a number of case studies. These case studies demonstrate how the Guides could be used by archaeologists to promote best practice in data creation and produce outputs that are suitable for long term archiving. The case studies will be used to illustrate the archiving of some of the specialist data types in archaeology, from creation and ingest through to dissemination. In this way, the real practical application of the Guides will be apparent.

The electronic nature of the new guides allows an integrated approach to these case studies. From each guide we will be able to link through to an exemplar archive which will illustrate the workflows that have been followed for the numerous data types and demonstrate what the final archived dataset might look like. This will be of particular value to archaeologists who are actively producing data, allowing them think about how their own data might look to other researchers once their fieldwork is complete.

All case studies are drawn from real and current projects in the academic and commercial worlds of archaeology.

8. TRENT-SOAR RIVER CONFLUENCE

The first of the case studies that we are working on is a study of the landscape surrounding the confluence of the Trent and Soar rivers in the East Midlands, England carried out by Birmingham Archaeology. Previous archaeological work on British river floodplains has suggested that river confluences can provide a focus for human activity through the ages. The distribution of archaeological remains in these regions is closely linked to the configuration of the landscape within the floodplain, both in terms of the original locations of sites and the level of preservation of the physical remains today. Attempting to accurately record and map this landscape was therefore a key element of this project [5]

In order to achieve this aim a number of different techniques and technologies were employed by the project team – aerial photography, lidar, geophysics (including GPR), GPS survey and GIS. This diverse and complex dataset is relevant to several of the Guides to Good Practice and can serve both to test the guidance and illustrate best practice in creating and submitting data that is suitable for long term archiving.

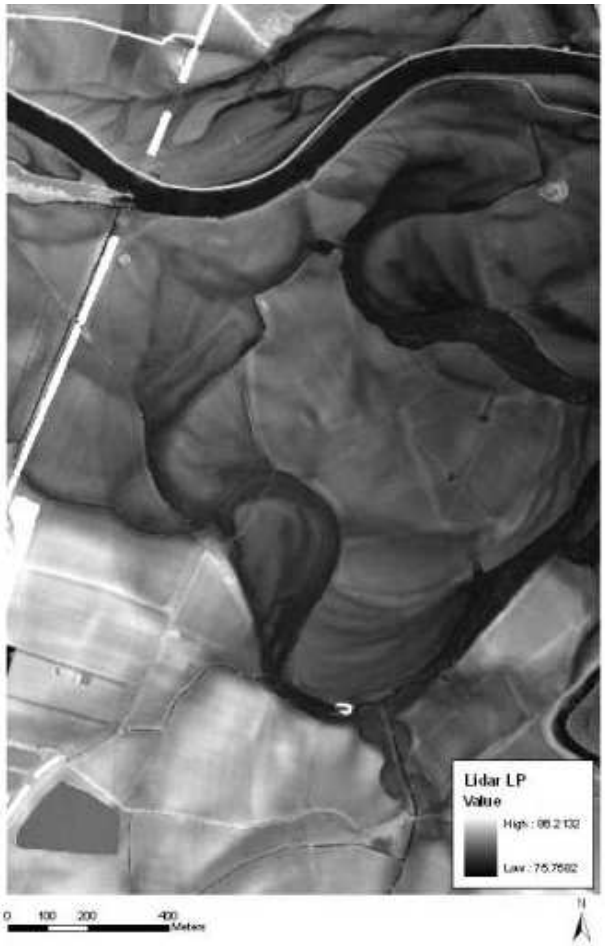


Figure 4. Lidar last-pulse (LP) surface model of the Trent-Soar confluence. Image © University of Birmingham. Lidar data © Infoterra Global Ltd

9. THE FUTURE

We have been working with our data producers for many years now, trying to ensure that the SIP we receive from them is adequate in terms of the types of files they send and the level of metadata attached to it. This however has never been a particularly easy job. We need to encourage data producers to think about digital

[5]See the following resource for phase I and phase II reports from this project which have been archived by the ADS http://ads.ahds.ac.uk/catalogue/resources.html?trentsoar_eh_2008

archiving from the very earliest stage of their project in order to ensure that they create their data in the right way with the right documentation. Alongside systems we already have in place such as on-line guidelines for depositors and metadata templates[6], we will soon be able to point people to these new 'Guides to Good Practice' from the outset of a project. The net result being a better, more complete SIP, and data that is 'independently understandable' to our designated community.

10. REFERENCES

[1] Consultative Committee for Space Data Systems *Reference Model for an Open Archival Information System (OAIS)*. CCSDS 650.0-B-1 Blue Book, 2002.

[2] Díaz-Andreu, M., Brooke, C., Rainsbury, M., & Rosser, N. "The spiral that vanished: the application of non-contact recording techniques to an elusive rock art motif at Castlerigg stone circle in Cumbria", *Journal of Archaeological Science*, Volume 33, Issue 11, pages 1580-1587, November 2006.

11. ACKNOWLEDGMENTS

The authors would like to acknowledge Keith Kintigh and Francis McManamon at Digital Antiquity, Arizona State University and Fred Limp at the University of Arkansas for their work on the Guides to Good Practice project. We would also like to thank the following funding bodies: Andrew W Mellon Foundation, English Heritage, and the European Commission VENUS project.

[6] http://ads.ahds.ac.uk/project/userinfo/deposit.cfm

PROPOSING A FRAMEWORK AND A VISUAL TOOL FOR ANALYSING GAPS IN DIGITAL PRESERVATION PRACTICE – A CASE STUDY AMONG SCIENTIFIC LIBRARIES IN EUROPE

Moritz Gomm
FernUniversität Hagen
Universitätsstrasse 1 58097
Hagen, Germany

Sabine Schrimpf
Deutsche Nationalbibliothek
Adickesallee 1
60322 Frankfurt/Main
Germany

Björn Werkmann
FernUniversität Hagen
Universitätsstrasse 1
58097 Hagen, Germany

Holger Brocks
FernUniversität Hagen
Universitätsstrasse 1
58097 Hagen, Germany

Matthias Hemmje
FernUniversität Hagen
Universitätsstrasse 1
58097 Hagen, Germany

ABSTRACT

In this paper we present a case study and selected results from a research on digital preservation amongst digital libraries in Europe. We propose a framework for gap analysis in digital preservation encompassing the diffusion of preservation practices and the life-cycle of data. We also present a Gap Analysis Tool that we developed to support visual analysis of gaps in the implementation of digital preservation amongst communities. We discuss selected results from the application of the tool in the community of libraries in Europe.

The authors would like to thank Eefke Smit from STM, Jeffrey van der Hoeven and Tom Kuipers from KB, and the four unknown reviewers for their valuable input and feedback. The research presented here was co-funded by the EC (Project PARSE.Insight, FP7-2007-223758).

1. INTRODUCTION

The survey results from the PARSE.Insight Community insight study [6] reveal the status-quo in long-term preservation of digital data in a variety of countries and institutions. The Gap Analysis Framework uses the survey data and matches it against framework elements for supporting the identification and interpretation of gaps between the current situation and what is necessary to enable secure long-term preservation of digital assets, with respect to particular groups of stakeholders. The Gap Analysis Tool (GAT) enables users (domain and preservation experts) to interactively visualize the results of the analysis and allows them to carry out more specific investigations into the highlighted gaps.

1.1. Gap Analysis Framework

We developed a Gap Analysis Framework that encompasses the life-cycle of scientific data (creation and use of data, re-use, preservation, and publishing) and the diffusion of digital preservation within scientific communities (awareness, knowledge, implementation, and commitment). The two orthogonal dimensions form the Gap Analysis Framework and are visualised in Figure 1.

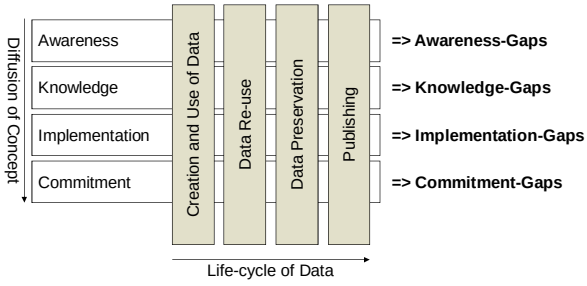


Figure 1. Gap Analysis Framework for digital preservation

While the four phases in the "life-cycle of data" are self explaining, the four aspects of the "diffusion of concept" need to be defined here:

- **Awareness** is the ability to perceive or to be conscious of the problems of long-term preservation in general.
- **Knowledge** is the sum of expertise and skills for the theoretical and practical understanding of long-term preservation issues. This includes knowledge about facts, information and means of long-term preservation.
- **Implementation** is the practical realization of means of long-term preservation including procedures, processes, systems and tools.
- **Commitment** is the willingness or pledge to preserve data.

1.2. Gap Analysis Tool

To support the application of the framework and to analyse the gaps in preservation practices we developed a tool within the EU-funded project "PARSE.Insight" [7]. The Gap Analysis Tool (GAT) analyses survey questionnaires used to gather information on preservation issues and calculate gaps in terms of the framework.

To allow for progressive refinement of search parameters and interactive data analysis [3], dynamic queries [10] and tight coupling [1] information visualization techniques have been employed. This way, immediate feedback is provided to enable interactive data analysis and visual scanning, to narrow down the choice of relevant information objects for a subsequent drill-down.

The drill-down metaphor is based on a "tree visualization" that employs regular expand and collapse operations as well as degree-of-interest [4] based pruning of the tree, to ease navigation. This allows for access to the information domain starting on a category level (e.g. the "awareness dimension" of the framework), down to the level of actual data items (e.g. answers to the survey question "do you have a preservation policy in place").

The user interface of the tool shown in Figure 3 is divided into two areas: The "Filter Setting" to the left, and the "Analysis View" to the right showing the described tree visualization.

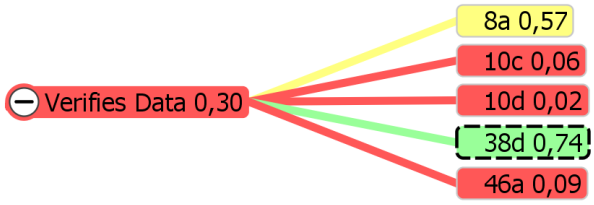


Figure 2. Gap Analysis Tree Detail: Leaf nodes to the right and the containing category to the left

Figure 2 shows an enlarged subsection of the tree with five leaf nodes and the containing category node. Leaves with dashed outline represent survey answers that indicate a problem (or gap), when chosen by a

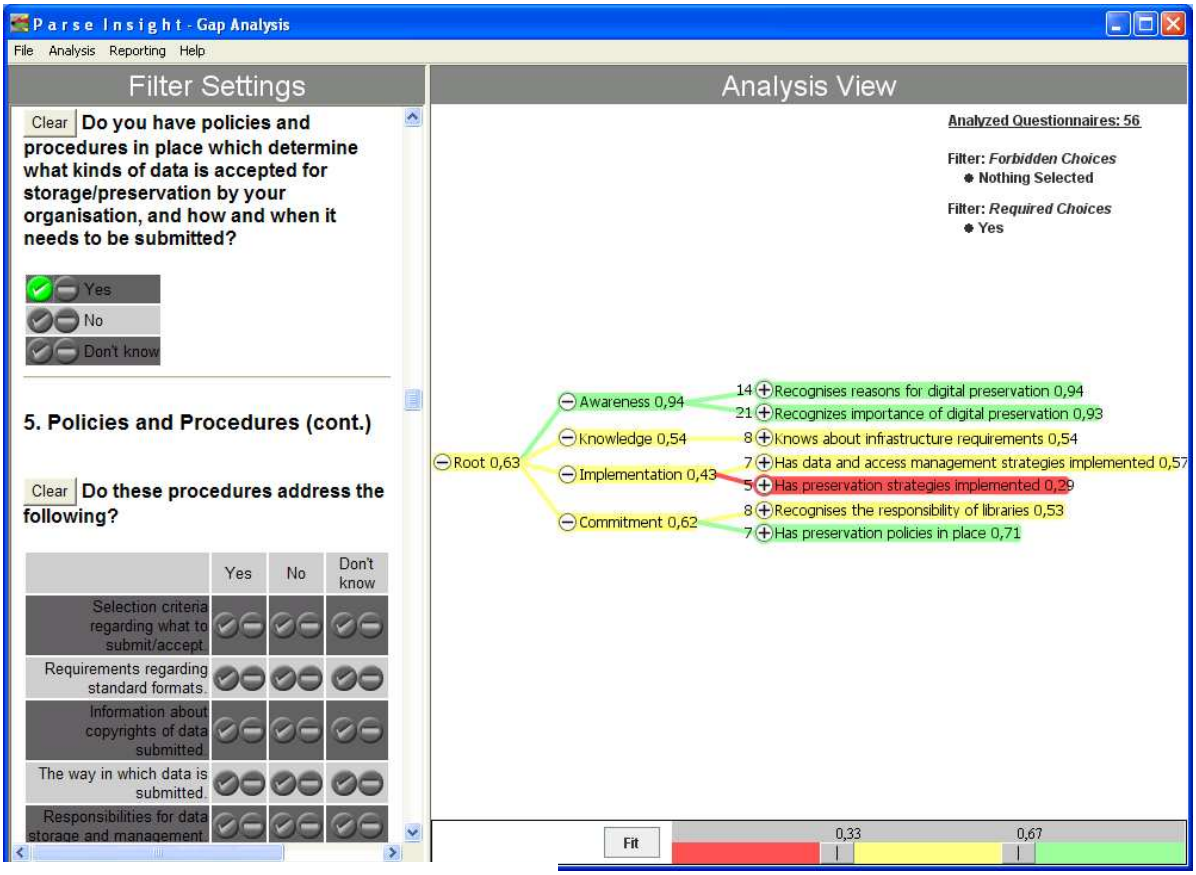


Figure 3. User Interface of the Gap Analysis Tool

participant. Answers depicted with solid outline indicate that no gap exists. Each leaf node has a label like "8a 0.57", giving the name of the represented answer, followed by a computed gap value. The *leaf gap value (lgv*) is computed according to the following formula:

$$lgv = \begin{cases} leaf\ indicates\ gap, & \frac{(pc - ac)}{pc} \\ otherwise, & \frac{ac}{pc} \end{cases} \quad (1)$$

where *pc*, the *participant count*, is the total number of participants of the survey, and *ac*, the *answer count*, is the number of participants that gave this answer. Hence, a high gap value, close to 1, is a good sign, while a lower gap value, close to 0, is indicative of problems within the set of analyzed participants. The set of analyzed participants, and consequently the *answer count* may vary depending on the filter settings as described later.

The gap values computed for the leaf nodes are propagated towards the root node according to the following formula for *node gap values (ngv)*:

$$ngv = \sum_{\substack{children \\ i\ of\ node}} i.gapValue * i.weight \quad (2)$$

The weights are chosen to sum to 1 to produce the average of the node values of the children. This was the case for all results presented here.

The nodes and edges of the tree are colored based on the gap values and the settings of a color slider. The color slider depicted in Figure 4 image for example, maps values from 0 to 0.23 onto the color red, while values between 0.65 and 1 will be shown in green. Intermediate values will be shown in yellow. This reflects the meaning of the gap value by showing gaps in red, as is also visible from Figure 2.

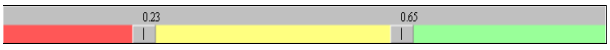


Figure 4. Slider

As mentioned above, the set of analysed participants is affected by the *filter settings* view, which shows a list of all survey questions and corresponding options for answers. Modifying these settings, it is possible to change the *answer count*, i.e., the subset of participants that are taken into account for the analysis view. In the beginning all participants are included. By clicking on the check mark button that is associated with every possible answer, the analysis view is configured to take only participants into account that chose the associated answer. This way, certain groups among the participant, e.g. those storing only certain types of data, can be analyzed separately.

2. CASE STUDY

2.1. Intoduction

Our tool was applied to a variety of communities. Here we want to present the case study on a survey among Libraries using data from LIBER (Ligue des Bibliothèques Européennes de Recherche). LIBER is the main research libraries network in Europe and encompasses more than 400 national, university and other libraries in 45 countries [5]. The survey was conducted as part of the PARSE.Insight Community insight study [8].

The tree structure for the survey data was modelled by the German National Library (Deutsche Nationalbibliothek, DNB) – a scientific library – and reviewed by individual LIBER members.

In the following we first present assumptions that were drawn from the LIBER-survey using classical empirical analysis methods before applying the Gap Analysis Tool. The results of our analysis were reviewed by individual LIBER members. Review was conducted on a voluntary basis, preceded by a call for review from the LIBER secretariat to the LIBER Working Group on Preservation and Digital Curation.

2.2. Assumptions from survey results

The following results from the survey attracted attention:

a) The great majority of the LIBER libraries recognize the reasons for and the importance of digital preservation. **Awareness** seems to be high.
b) The majority of the participating libraries believes that an international infrastructure would help to guard against the threats of digital preservation (66 %). Furthermore, the majority of libraries is convinced that more is needed for digital preservation, above all more resources, more knowledge, more digital repositories, and more training opportunities. **Knowledge** about digital preservation requirements seems to be high, too.
c) The majority of libraries claim that they do already have policies and an infrastructure in place (59 %). However, only 27 % believe that the tools and the infrastructure available to them is sufficient for their digital preservation objectives, as opposed to 56 % who believe not so. There seems to be an **implementation** gap.
d) The majority of the libraries consider National libraries and research libraries responsible for digital preservation. Additionally, for about 75% of the participating libraries, funding for digital preservation is and will also in the future be an issue. This shows that there is a lot of **commitment**.

The gap analysis tool then was used by the DNB staff to check these assumptions and to render some findings more precisely.

2.3. Preparation of the Gap Analysis Tool

A total of 70 items were identified and grouped according to the framework (see Table 1). The selection of question items and grouping into categories was subject of the review by LIBER members.

Dimension	Sub-categories for survey items
Awareness	Recognises reasons for digital preservation Recognises importance of digital preservation
Knowledge	Knows about infrastructure requirements
Implementation	Has data/access management strategies Has preservation strategies implemented
Commitment	Recognises the responsibility of libraries Has preservation policies in place

Table 1. Sub-categories from the LIBER-survey

2.4. Gap Analysis of the LIBER data

The visualization of the base data gives a slightly different picture from the assumptions above (see Figure 5): Only awareness is – as assumed – marked with a positive gap value (green colour), while commitment is on a modest level (yellow) and knowledge is even low (red). The implementation gap that was assumed can be confirmed.

Figure 5. Visualization of the libraries survey data

The experts now analysed the data further by drilling in and out and selecting different subsets of the data which caught their interest. For example if they want to find out the differences in the community between those who have appropriate "policies and procedures" in place compared to those, that haven´t, they only need to change the filter setting for the corresponding question and compare the visual results (see Figure 6). Other filters that were set include for example:

- The kind of data stored in organisations (data-sets, e-books, or e-journals)
- The volume of data stored in organisations
- Preservation strategy in place
- The kind of preservation strategy in place (migration, emulation, or outsourced to third party)
- Confidence that the organisation's infrastructure will scale with future requirements
- Opinion what is needed to guarantee reliable preservation measures (we distinguished between training, more resources, more repositories/archives)

The time and effort required for the entire analysis was about one personal month plus a few hours of technical support. The external reviews of the results took approximately half a day per reviewer. It should be noted, that the effort was relatively high because feedback from the experts was also used to further improve the Gap Analysis Tool.

2.5. Findings

2.5.1. Policies and Infrastructure

A clear relation between selection policies and the level of implementation and commitment could be shown. Libraries that have thought of what kind of content they add to their collections and documented that in writing in their selection policies are more committed to the task of digital preservation and are better prepared in terms of implementation – although there remains a gap in terms of implemented preservation strategies.

What seems to be important is the fact that there are selection policies in place. The kind of material, however, that libraries collect does not seem to have a heavy impact on libraries' preparedness for digital preservation. No matter if they are focussing on more traditional publication types like e-books and journals or on for libraries unfamiliar data sets – the gaps remains almost the same.

2.5.2. Amount of Data

In contrast, the amount of data that a library currently stores seems to have an impact on the gaps in preservation. The larger the amount of data that a library has to deal with, the smaller the gaps in the area of implementation and commitment are. There is a direct relation between the fact that a library stores data and feels responsible. Another relation could be shown between the amount of data and the implementation of data and access management strategies.

2.5.3. Preservation Strategies

Since a gap is indicated in the area of implementation of preservation strategies in all analyses it is instructive to look in more detail at those institutions that have already implemented preservation strategies in comparison with those that have not implemented the respective strategies.

There is a "commitment gap" in the category "recognises the responsibility of libraries" which cannot be explained easily. The gaps get even bigger when we look at those institutions that state explicitly that they do not have preservation strategies in place. Here awareness remains high, but the values in knowledge, implementation and commitment are significantly lower in comparison to the institutions that have strategies

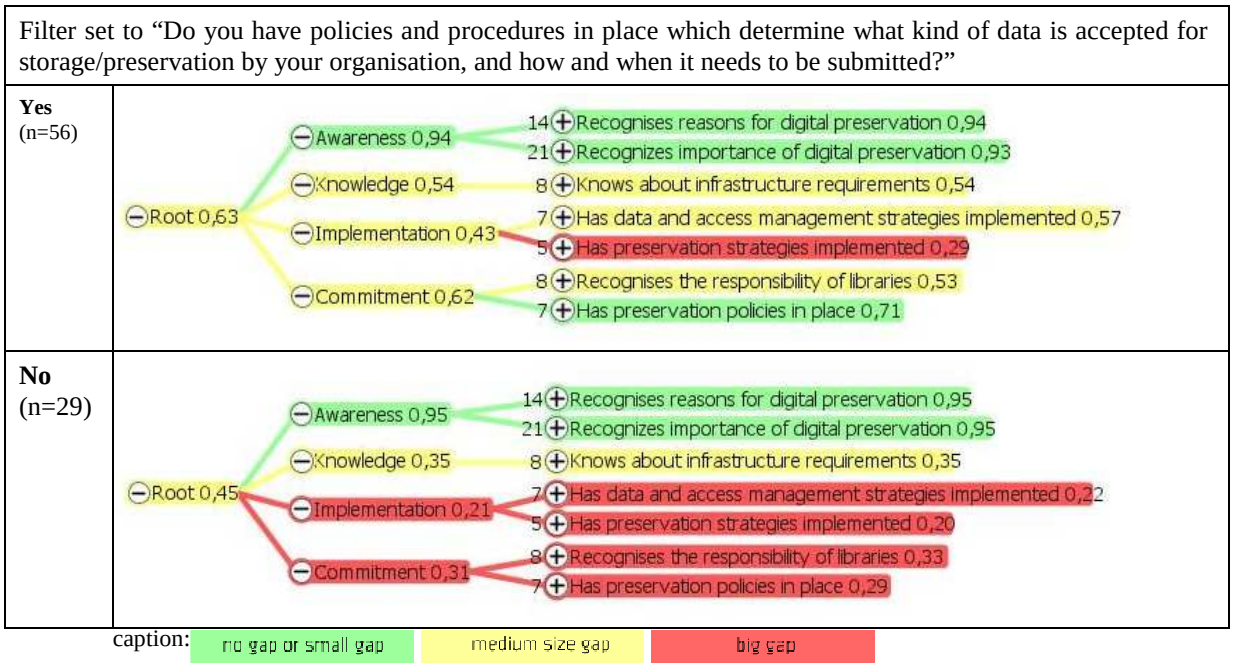


Figure 6. Example of visual analysis: comparing two groups of respondents

implemented. The relation is obvious: Implementation of preservation strategies requires knowledge, results in implementation and facilitates commitment.

What should concern the library community is the fact that the institutions without implemented strategies are far behind in most categories. In order to catch up, they need to start with building up knowledge.

2.5.4. Scalability

When we compared those libraries that are confident that their infrastructure will scale with future requirement with those that are not so confident, we find a distinction mainly in the areas of knowledge and commitment. Again, the area of preservation strategies catches the eye. While there is only a small gap at those institutions that feel prepared, it is the largest gap at those institutions that feel not prepared for future requirements.

2.6. Summary

The framework and the gap analysis tool allowed deeper insight into the gaps within the scientific libraries community and showed some relation between gaps that were not obvious before.

The first visualization of results indicated larger gaps than could be expected from a simple review of the survey results. It must be acknowledged, though, that many survey participants had skipped many answers that were not mandatory, while skipped answers were counted as negative answers. For future study designs that make use of the Gap Analysis Tool we will take this finding into account and exclude optional questions as far as possible from the surveys.

However, the gap analysis with the tool proved the assumption right that there is mainly an implementation gap, which can be explained with a gap in the implementation of preservation strategies. The gap analysis furthermore indicated a relation between missing preservation strategies and little knowledge and commitment within the respective libraries.

The results also indicated that there is a difference between large and small archiving facilities: The more data a library has to store, the lesser its gaps in the areas of knowledge, implementation and commitment are, hence the better it is prepared for digital preservation. The results indicate in a similar way that libraries with preservation and selection policies in place have smaller preservation gaps than those who have not. The largest difference is between those libraries that have or have not implemented preservation strategies.

Overall, the results indicate a gap between well prepared and less prepared libraries. The less prepared libraries must be attentive that they do not fall behind. Means to close these gaps are discussed in the PARSE.Insight Roadmap [9].

In general, the Gap Analysis Framework and Tool can be used for assessing current preservation practices and benchmarking progress within or compare results between given communities of practice. Of course the basic prerequisite is the availability of sufficient survey data on which the gap analysis can be based.

The analysis indicated that the four aspects of the framework dimension regarding the "diffusion of concept of digital preservation" are not fully independent of each other. From our research results they seem to be interrelated as follows:

a) Implementation requires basic knowledge
b) Knowledge hardly exists without awareness.
c) Commitment requires awareness and can be strengthened by knowledge
d) Commitment can exist without implementation which then is considered a "lip service"
e) Systems can be in implemented without being used, if the commitment of using them is missing.
f) Commitment can be found on a corporate level (e.g. policies) and on a personal level (willingness to use the implemented systems)

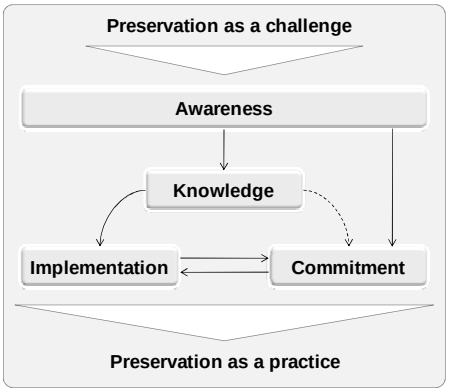


Figure 1. Relations and dependencies of the aspects of the diffusion of the concept of digital preservation

In further research projects we will refine the Tool and investigate how it can be integrated with other tools such as the AIDA-Toolkit (Assessing Institutional Digital Assets) [2] for analysis of institutional levels.

3. REFERENCES

[1] Ahlberg, Chr. and Shneiderman, B. (1994): *Visual Information Seeking: Tight coupling of dynamic query filters with starfield displays*, Proc. of ACM CHI94 Conference (April 1994), 313-317 + color plates.

[2] AIDA Self-Assessment Toolkit, URL: http://aida.jiscinvolve.org/wp/

[3] Card, S., Mackinlay, J., and Shneiderman, B. (1999): *Readings in Information Visualization: Using Vision to Think.* Morgan-Kaufmann.

[4] G.W. Furnas, *Generalized fisheye views, SIGCHI Bull.*, vol. 17, 1986, pp. 16-23.

[5] Ligue des Bibliothèques Européennes de Recherche (LIBER), Website: http://www.libereurope.eu.

[6] PARSE.Insight: *Insight Report. Insight into digital preservation of research output in Europe, 2010. URL: http://www.parse-insight.eu/downloads/ PARSE-Insight_D3-6_InsightReport.pdf*

[7] PARSE.Insight: *Insight into issues of Permanent Access to the Records of Science in Europe.* Project No. 223758. (EU-funded Project)

[8] PARSE.Insight: *survey results*, URL: https://www.swivel.com/people/ 1015959-PARSE-insight/group_assets/public

[9] PARSE.Insight: *Science Data Infrastructure Roadmap, 2010. URL: http://www.parse-insight.eu/downloads/PARSE-Insight_D2-2_Roadmap.pdf*

[10] Shneiderman, B., Williamson, Chr., and Ahlberg, Chr. : "Dynamic Queries: DataBase Searching by Direct Manipulation", In *Proc. of Human Factors in Computing Systems*, CHI '92, ACM Press, 1992, pp. 669-670.

'DIGITAL PRESERVATION: THE PLANETS WAY': OUTREACH AND TRAINING FOR DIGITAL PRESERVATION USING PLANETS TOOLS AND SERVICES

Vittore Casarosa **Laura Molloy** **Kellie Snow**

Humanities Advanced Technology and Information Institute (HATII)
University of Glasgow
11 University Gardens, Glasgow G12 8QJ

casarosa@isti.cnr.it l.molloy@hatii.arts.gla.ac.uk k.snow@hatii.arts.gla.ac.uk

ABSTRACT

This paper outlines the Europe-wide programme of outreach and training events, jointly organised by HATII at the University of Glasgow and the British Library, in collaboration with a number of European partner institutions, on behalf of the Planets project (Preservation and Long Term Access Through Networked Services) between June 2009 and April 2010. It describes the background to the programme and the events which took place during the final year of the project, focussing on the success of the events based on feedback results, lessons learned from the production of the series, and the perceived long-term impact of the programme on future Planets and digital preservation training activities.

1. THE PLANETS CONTEXT

The Planets project was a four-year project co-funded by the European Union under the Sixth Framework Programme to address core digital preservation challenges. It ran for four years from 1 June 2006, delivering research, tools and services resulting from the collaboration of sixteen partners across Europe, including national archives and libraries, higher education and research institutions and major IT companies.

The main aim of the project was to design, build and deliver practical tools and services to enable long-term access to cultural and scientific digital assets across Europe. These tools and services were planned to be highly automated and easily scalable, in order to minimise costs and maximise compatibility with the widest possible range of users. Main achievements of the project include the development of Plato preservation planning tool, the Testbed experimental preservation environment, and an extensive range of innovative research into a variety of digital preservation issues, methodologies and approaches to help users understand, define, and protect their collections, and to approach digital preservation in an integrated way.

2. THE PLANETS TRAINING APPROACH

As part of its remit to maximise efficient uptake of Planets tools and services, the project delivered a training programme offering learning opportunities to staff in memory institutions such as national archives, libraries and large content-holding organisations. The perceived role of the training was not to organise and deliver events in isolation; instead it was expected to be a public face of the project, providing a conduit between the innovations of the project and its user communities to maximise take-up of Planets methods, products and services. As a result, a comprehensive and timely programme was required which would fulfil the needs and requirements of a number of groups.

Work on the programme began with the production of a detailed training plan[1]. The plan looked at developing a programme using a modular approach, which would provide self-contained sessions which could be integrated with other projects at collaborative events, as well as combined for longer Planets-specific events as tools and services were finalised.

A provisional programme of events was outlined for the duration of the project, alongside a significant amount of detail on how the event activities would be documented and evaluated. The planned programme adopted the approach of initially delivering short Planets sessions as part of more general digital preservation training events, whilst Planets tools and services were still in development, in order to educate institutions throughout Europe in the general principles of digital preservation and to raise awareness of the project. More detailed Planets events would then be organised towards the end of the project. This would allow the research

[1] http://www.planets-project.eu/docs/reports/Planets_DT6_training_plan_final.pdf

outputs and the tools developed by the project to be presented to interested members of the public at a point where these outputs had reached a relatively stable stage of development when their value could not only be described but also demonstrated.

The project proceeded to deliver a number of joint events during years two and three of the project, as part of collaboration with the FP6 Digital Preservation Europe ('DPE') and Cultural, Artistic and Scientific knowledge for Preservation, Access and Retrieval ('CASPAR') projects, and including training on Planets and its initial results. As the project entered its final year, the original training plan was then refined and extended[2] to offer a series of Planets-specific outreach and training events with supporting online materials.

3. EARLY ACTIVITIES – THE COLLABORATIVE APPROACH

The aim of the early training activities was not only to deliver training events on the Planets approach, but also to embed the training into the wider digital preservation education initiatives of other FP6 and European digital preservation projects.

With the establishment of the wePreserve[3] consortium, initiated by the DPE, Planets and CASPAR FP6 projects, Planets took the opportunity to collaborate on the delivery of a series of introductory events which covered the general aspects of digital preservation. Courses were delivered in Vilnius (October 2007), Prague (October 2008) and Barcelona (March 2009). Using its modular approach Planets was able to insert sessions about developing Planets tools into the overall programme, offering a much more appropriate form of training in the early stages of the project as tools were emerging at different stages.

The three projects also established corresponding Virtual Learning Environments (using Moodle software) which were used to provide additional pre-course training materials for attendees. After the Vilnius event, materials were displayed on one Moodle (that of the DPE project) only, in order to avoid confusion for delegates and further combine the approach to the training events.

The strategic and co-ordinated approach to early events enabled the Planets training programme to have a much wider impact on digital preservation training for the European Community. Rather than potentially restricting take-up through focusing on project-specific training for the duration of the project, Planets tools and services were instead introduced to more diverse audiences in a general context that made training more palatable and in turn fostered interest in the Planets approach. At the same time a collaborative approach to events did provide drawbacks; the designing of programmes that fulfilled each project's requirements was a continued challenge and prolonged the planning process significantly. The degree to which each project's results could be fully disseminated in a joint event was also limited, and as each project matured this became a more obvious issue. The original anticipated need for detailed project-specific training was therefore confirmed as Planets moved towards completion.

4. THE 2009-2010 TRAINING EVENTS AND ACTIVITIES

Whilst the collaborative training events proved successful, it was equally clear that there was a need, particularly amongst the growing Planets User Community, for more courses dedicated to the Planets approach. Outreach and training efforts within Planets had heretofore been viewed as discrete resources, but after the delivery of a highly successful combined outreach and training event, focusing on preservation planning using Planets tools, in Vienna in April 2008, it was agreed that effort between the Planets training and outreach services should be combined for a final series of events dedicated to introducing the now more complete Planets approach to the core target audiences.
Five outreach and training events were planned during the fourth and final year of the Planets project. Locations for the 2009-10 events were carefully chosen in order to reach as broad a range of European countries and contexts as possible, and the events were publicised through the extensive network of digital preservation, archive and library-related mailing lists across Europe. Each event placed a target on attracting at least seventy percent of participants from the local region, and promoted the event through regional contacts and organizations.

The first event took place in Copenhagen, Denmark in June 2009. This was followed by Sofia, Bulgaria (September 2009), then Bern, Switzerland (November 2009), London, England (February 2010) and finally Rome, Italy, in April 2010. The training team carried out initial research into the level of digital preservation activities within each region of Europe, in order to tailor courses to their anticipated audience. A pre-event questionnaire was then distributed to delegates prior to the event and the results disseminated to speakers, to ensure individual sessions were pitched correctly. The work identified that knowledge and activities in Southern and South Eastern Europe were less advanced than those in Northern and Western Europe, and so the Sofia and Rome events had a different regional focus to those of Copenhagen, Bern and London.

Each event consisted of an initial day of high-level explanation of the challenges of effective digital preservation, along with an overview of Planets

[2] http://www.planets-project.eu/docs/reports/Planets_DT6-D4_Training_Plan.pdf
[3] http://www.wepreserve.eu/

solutions to these challenges: this initial day was targeted at managers, budget-holders, policy-makers and other senior decision-making members of staff.

Days two and three consisted of a mixture of lectures on more detailed technical information about the Planets tools and services, interspersed with practical demonstrations of the tools working live and opportunities for open discussion. These two days were aimed at librarians, archivists, and the technical and developer staff who would be involved in the implementation and maintenance of Planets tools, should they be adopted by their institution.

Each event also incorporated one or two guest speakers who gave a more personal account of either region-specific digital preservation concerns, or a case study of how they had tackled the digital preservation issues, very often using Planets tools and services within their institution. Speakers from institutions such as the Central European Library, Bavarian State Library, Bibliotheque Nationale de France and UK Parliamentary Archives discussed their experiences alongside Planets partners from the National Library of The Netherlands and the Swiss Federal Archives.

Alongside the training events, HATII also led the development of a suite of online training materials, both to complement the learning of those who had attended one of the year four courses, and also to introduce the principal concepts of the Planets approach to digital preservation to anyone unable to attend the face-to-face training. The use of materials on the Moodle sites for joint events had been lower than anticipated, with feedback suggesting neither delegates or tutors had sufficient knowledge of the software or time to familiarise themselves with how to use it correctly. The Copenhagen event trialled placing supporting materials on a Planets webpage, which was well received, and as a result the decision was made to use a dedicated area of the Planets website for dissemination of final online training materials instead.

These materials were made freely available online at the close of the fifth and final 'Planets Way' event, and consist of seven short videos, based on day one presentations, plus an annotated reading list and a set of summaries of the first day outreach material prepared by IBM (one of the Planets industry partners) for technical and development staff. This material is currently hosted on the Planets website, though this may change due to future activities of the Open Planets Foundation (OPF). This not-for-profit organisation has been established at the end of the project in order to continue the development and support of the Planets tools including training for interested organisations.

5. FEEDBACK

After each 'Digital Preservation – the Planets Way' event, delegates were asked to complete a feedback form, scoring various aspects of the course and also providing comments on what they liked best about the event and what could have been done better. The delegate feedback from each event was carefully gathered, monitored and integrated into planning for the next event in the series.

The events were well attended, with three out of five exceeding attendance target for all days, namely fifty delegates for day one and thirty for days two and three. The regional focus for events was also successful, with four out of five reaching the target of seventy percent of attendees being from local countries. A good proportion of attendees were however from a range of countries outside Europe, including the USA, Australia, and Saudi Arabia. Despite efforts to attract delegates from all regions of Europe the countries of Southern and South Eastern Europe were still underrepresented across the series, reflecting the lower level of involvement in digital preservation activities or perhaps a reluctance to attend events delivered in English.

Delegates represented a wide variety of national libraries, national archives, academic and government institutions, and within these organisations the target job functions of librarians, archivists, CEOs and IT staff were well represented, alongside digital preservation researchers.

The feedback to a set of event and organisational criteria was consistently high across the events (Figure 1), with many areas improving as the series developed and trainers listened to the comments of attendees. A target was set for seventy percent of delegates to rate events as good or better against criteria. The courses' success in providing a good introduction to Planets and meeting expectations were particularly highly rated and a significant majority would consider using Planets and attending similar events in the future. The areas which received the lowest average scores were enabling delegates to understand approaches and the gaining of practical skills, but even these remained close to target. The various organisational aspects of the events also received excellent feedback, with the speakers and content of sessions highly praised, and pre-event reading and the content of exercises generally performing less well.

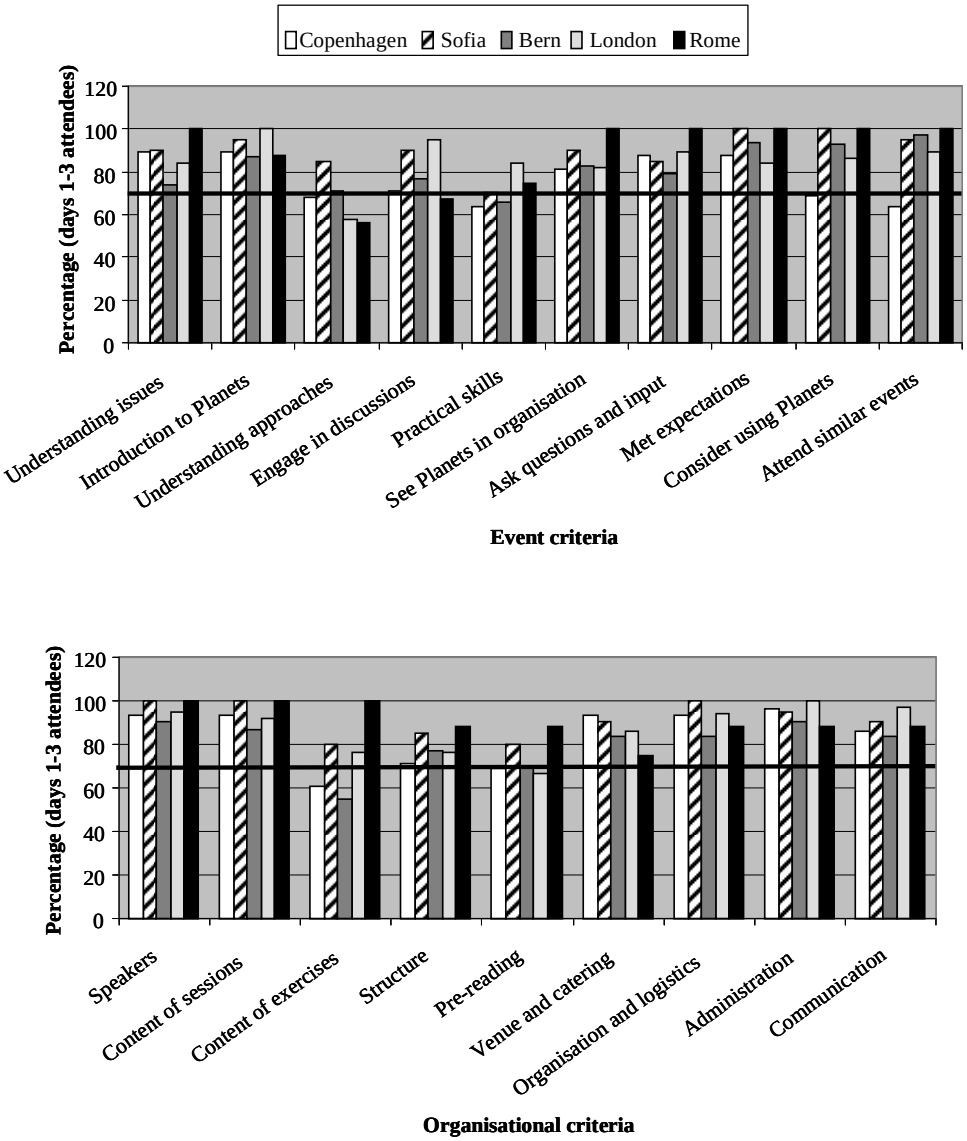


Figure 1. Feedback across the series.

In the feedback form, participants were asked to rate each lecture individually, and the ratings and comments were carefully examined by speakers in order to improve individual sessions. For the overall average rating of all the lectures in each event (Table 1), excluding Sofia the rating was constantly improved, reaching very high levels.

	Avg rating	Std deviation
Copenhagen	4,05	0,32
Sofia	4,47	0,16
Bern	4,09	0,27
London	4,23	0,17
Rome	4,25	0,17

Table 1. Average rating for lectures.

The standard deviation, which indicates how much on average the individual lectures are above or below the overall value, decreased from the first to the last event, indicating that not only did each lecture improve over time, but also the difference in rating between each lecture gradually reduced.

In addition to scoring various criteria for the training events, delegates were invited to provide general comments on what they liked best about the event, what they would like to see in future events and what they thought could have been done better. General consensus across the series was that the practical exercises were a favourite part of the events, as well as the opportunity to network with other attendees and the knowledgeable speakers. Delegates would have liked to have seen more examples of Planets being used in an institutional

setting, a clearer explanation of how the tools fitted together, and more opportunity to use tools individually during the exercises.

The comments on what could have been done better highlighted the difficulties of satisfying audiences with a variety of job roles, institutional contexts and digital preservation knowledge. Delegates requested differing levels of introductory information on digital preservation, and whilst some commented the technical level of the event was too low, librarians and archivists in particular tended to suggest it was too high. This conflict of opinion supported other comments which indicated that some areas of the events should distinguish between information needed for librarians, archivists and managers, and that required by IT professionals or developers. Several delegates also commented that the events tried to fit too much information into a three-day course, suggesting that courses tailored to more specific audiences may go further in addressing the specific needs of target groups.

Other methods of feedback were also used and valued by the organiser team, including spontaneous suggestions on the part of the audience which the organiser team used and carried on to further events. For example, the use of Twitter was initiated by a delegate at the first event in year four, in Copenhagen. This delegate created a hashtag ('#Planetsway') in order to identify Twitter messages specifically about the Planets event. In subsequent events of the series, the organiser team encouraged delegates in both the use of Twitter for feedback and that specific hashtag.

Blogs were also useful pieces of evidence after each event. At the end of each event, feedback from Twitter and the blogosphere was searched for and included in the evaluation process. Blogs that posted entries about the year four events include Archives Hub blog[1], the KeepIt project's 'Diary of a Repository Preservation Project'[2] and the Bodleian Library's FutureArch blog[3]

6. IMPACT

At the end of the series of 2009-10 outreach and training events, a post-event survey was conducted. Feedback from each event had routinely shaped the content and format of subsequent events throughout the training, but to truly judge the success of the training programme, it was important to gather information about the real impact the programme had had on its attendees' working practices and in turn the national library and archive and digital preservation communities of Europe.

The survey attempted to find out two things; firstly, the extent to which delegates had implemented the knowledge and skills they had gained on the course, and secondly, whether the events had led to any collaborations or increased involvement for delegates in digital preservation forums or Planets activities.

A questionnaire was designed with a set of yes/no questions asking for further description where required. The questionnaire was issued to all known attendees of the training courses across the programme (excluding Rome which had not taken place at the time of the survey), which totalled 339 individuals. A small prize was offered as an incentive for delegates to respond. Seventy delegates responded to the questionnaire (some partially), giving an overall response rate of twenty-one percent. The percentage of responses for each event against the number of attendees varied, with early events often as well represented as later courses.

Three questions were asked focusing on the implementation of knowledge and skills. The first asked whether since attending the course delegates had gained knowledge and/or skills which they had been able to implement in their work. Fifty replied that they had, with only four claiming they had not. Some responses commented that the event had helped them to understand the general issues surrounding digital preservation, whilst several specified migration and emulation, significant properties and preservation planning as skills which they had been able to implement. A number stated that nothing had yet been implemented as they were not at that stage within their organisation, but that the training had helped them understand how to approach the issue; as one delegate responded: "We are in the process of creating strategies for preservation and Planets has given me awareness about tools and services that I can use to achieve that".

The second question inquired whether as a result of attending the course delegates had introduced, or anticipated introducing, new activities/initiatives in their organisation to preserve digital content. Thirty-nine delegates concluded they had, specifying a variety of activities including attributing metadata and significant properties and general digitisation. One delegate commented that whilst existing initiatives in their organisation had stayed the same, "Planets does provide useful tips and methodologies to improve the effectiveness of those".

The third and final question in this category considered whether delegates' organisations had implemented, or intended to implement, any of the Planets tools and services. There were an almost equal number of positive and negative responses to this question. General consensus seemed to be that many were planning to but had not yet done so, and in some cases were just beginning initial testing with some of the tools. Out of twenty-seven positive responses, thirteen specifically named Plato and/or the Testbed as services they planned to implement.

[1] http://archiveshub.ac.uk/blog/?p=6

[2] http://blogs.ecs.soton.ac.uk/keepit/2010/02/16/planets-way-london-highlights/

[3] http://futurearchives.blogspot.com/2010/02/music-planets-and-secret-messages.html

The second part of the questionnaire asked five questions about delegates' involvement in the digital preservation community and with Planets in particular. The first asked if any collaborations or working relations had been established as a result of attending a Planets course. Twenty-six delegates specified they had, in particular citing continued contact and sometimes even collaboration with speakers, and partnerships or collaboration with other institutions who had attended the course. The second question asked if delegates had become involved in discussions about issues raised in the course through discussion lists or forums. Only thirteen responses stated that they had, listing discussions with colleagues on a local level as well as following discussions through mailing lists and forums. The responses to both these questions demonstrated the importance of the face-to-face aspect of the programme and its role in encouraging networking and in fostering the development of the digital preservation community.

The next two questions asked whether delegates had participated in any further digital preservation training or Planets dissemination activities since attending the course. A relatively low number of positive responses were received to both of these questions, suggesting it might be useful to place more emphasis on attracting previous delegates to any future events. Further training and dissemination activities listed included other Planets and digital preservation project events, various local workshops and the Planets community and newsletter.

The final question asked whether attendees or their organisation would consider subscribing to Planets technology or becoming a member of the Planets project follow-on organisation, the Open Planets Foundation (OPF). Forty-three respondents answered that they would consider this, with only ten specifying no. Comments indicated that in order to decide the advantages of joining would need to be weighed against the costs of subscription.

The post-event survey confirmed that the Planets training events had had a long-term impact on delegates' preservation activities within their organisations, in particular providing them with the skills necessary to tackle the issue of digital preservation and to implement new activities as a result. There was a continued interest in the Planets tools and services, with organisations already testing components and interested in the work of the OPF. The events were also significant in encouraging networking and long-term working relationships. Perhaps where the events had less impact was in encouraging delegates to become involved in other digital preservation and Planets activities. This suggests further effort should be dedicated to promoting future activities with previous attendees who will already have an interest in and understanding of the tools and services on offer.

7. CONCLUSIONS AND RECOMMENDATIONS

The Planets project delivered a successful training programme which reached a significant number of delegates throughout Europe and the rest of the world. The experiences and feedback from the programme offer a number of recommendations for future digital preservation training programmes which the project has identified:

Personalise courses for different occupations and geographical regions – the feedback from the events demonstrated that different audiences require different levels of training. Future training programmes should consider offering separate events for different occupation types and regional areas to ensure that the level of training is precisely suited to its audience.

Use alternative approaches to generate interest in countries less involved in digital preservation – despite efforts events were predominantly attended by countries already active in digital preservation. Further investigation into the most useful types of digital preservation training for regions underrepresented at events should be considered.

Encourage opportunities for collaborative training events – the joint training activities were an excellent way of raising awareness of what the project can offer to a broader digital preservation community. Many attendees have gone on to test and implement various tools within their own institution as a result of learning about early project developments.

Place an emphasis on practical sessions and real-life examples – delegates consistently praised the practical element of events and requested more hands-on activities and case studies to place theory into context.

Use face-to-face training events – attendees emphasised the opportunity to share ideas with other delegates and speakers in person as one of the highlights of their experience. Events support broader outreach activities and help to build a community receptive to subsequent project developments.

Develop effective online training facilities – the potential of online training is substantial as it is able to both support physical events and educate individuals unable to attend courses. The design of effective training tools however requires significant consideration and investment in order to ensure they are useful.

Use effective planning and evaluative procedures – the constant reassessment of the training programme ensured its success.

SUSTAINABILITY CASE STUDY: EXPLORING COMMUNITY-BASED BUSINESS MODELS FOR ARXIV

Oya Y. Rieger **Simeon Warner**

Cornell University Library
Ithaca, NY
USA

ABSTRACT

arXiv.org is internationally acknowledged as a pioneering and successful open access digital archive for research articles. The case study discusses the efforts to establish a community-based sustainability strategy to ensure the longevity, effectiveness, and success of the service. It also describes the costs associated with running the repository that take into consideration both daily operational costs and efforts in improving its technical architecture and functionality.

1. INTRODUCTION

Started in August 1991, arXiv.org emerged as an exemplary disciplinary digital archive and open-access distribution service for research articles. The e-print repository has transformed the scholarly communication infrastructure of multiple fields of physics and plays an increasingly prominent role in a unified set of global resources for physics, mathematics, computer science, and related disciplines. It is firmly embedded in the research workflows of these subject domains and has changed the way in which scholarly articles are shared, making science more democratic and allowing for the rapid dissemination of scientific findings.

arXiv moved to Cornell in 2001, and is operated and maintained by Cornell University Library. The Library is committed to maintaining arXiv as an open access service, free to submitters and users alike. However, we believe that as a public good, arXiv should be supported by those institutions that use it the most. In an effort to address the long-term sustainability of this critical open access repository, the Library has developed a collaborative business model based on income generated by contributions from the institutions that are the heaviest users of arXiv [1].

In the first part of this case study we will describe the business planning process to provide a case study of opportunities and impediments in developing alternative business models. Financial stability alone is not sufficient to sustain a service such as arXiv, it must also be developed and improved to meet the needs and expectations of those who use it. Also critical is understanding the long-term preservation challenges associated with running a repository. In the second part of the case study we will outline our evolving technology plans including the underlying platform, preservation, services and interoperability.

2. SUSTAINABILITY: PROVIDING ENDURING ACCESS

Ithaka's 2008 report on sustainability provides a comprehensive review of a variety of business models for supporting online academic resources [2]. The report defines sustainability as "the ability to generate or gain access to the resources financial or otherwise needed to protect and increase the value of the content or service for those who use it." Therefore, keeping open access academic resources such as arXiv sustainable involves not only covering the operational costs but also continuing to enhance their value based on the needs of the user community. Furthermore, sustainability involves running a robust technical operation and addressing the digital preservation requirements of a system to ensure its long-term longevity and usability. Such a financial commitment is likely to be beyond a single institution's resources.

Scholars worldwide depend upon the stable operation and continued development of arXiv. Sustainability is best assured by aligning revenue sources with the constituents that realize value from arXiv, and by reducing dependence upon on Cornell University Library's budget. Our collaborative business model aims to engage the institutions that benefit from arXiv in defining the future of the service.

3. BUSINESS MODEL

3.1. Business Planning Process and Motivating Factors

arXiv moved to Cornell in the summer of 2001, and the Cornell University Library currently provides the bulk of arXiv's operating costs. Currently at $400,000 per year, the cost of operating arXiv is comparable to the entire physics and astronomy collections budget of the

Cornell University Library. The business planning process began in June 2009 necessitated by the significant budget cuts that required the Library to review its programs and funding sources. Although the formal sustainability planning process was triggered by financial pressures, the Library was already engaged in exploring funding sources to support and further develop the archive, such as seeking an endowment and considering grant funding opportunities.

The first phase of the sustainability planning process involved a landscape analysis and a survey of arXiv stakeholders' positions and opinions on the e-print repository's future. Also critical during the assessment phase was expanding our understanding of the income models for open access and pros and cons of emerging practices. As we surveyed the positions of administrators, managers, and scholars from libraries and research centers, we often received the following concerns and questions:

- When is arXiv going to replace the formal journals?
- How will you address the free rider problem?
- Why not charge scholars per submission?
- What are the benefits for my institution?
- How will you structure a governance model?
- Are you opening a floodgate? Will other open access initiatives also start requiring contributions from the user community?
- What are the other potential sources of revenue?
- What is your long-term plan?

Based on a thorough review of available funding models [2, 3] and an extensive survey of arXiv stakeholders, we have considered many possible support options that are compatible with the Cornell University Library's mission. These include: sponsorship and advertising; donations; endowment; creation of "freemium" services; and support from funding bodies, scholarly and professional societies, and publishers. We consider the current plan as short-term and over the next three years will work with our advisors and supporting institutions to develop a long-term plan. The arXiv white paper further describes our planning process as well as addressing the questions raised by stakeholders during the input gathering process [1].

We also have considered the role of the Sponsoring Consortium for Open Access Publishing in Particle Physics (SCOAP3) initiative for our financial planning.[1] arXiv would potentially be a beneficiary of redirected funding administered by the SCOAP3 consortium. It is not clear, however, when this initiative will meet its funding goal. It should also be noted that SCOAP3 is restricted to High Energy Physics (HEP) and particle physics content only, which represents between 18% and 40% of submissions to arXiv (depending on how broadly the subject area is construed). Thus SCOAP3 could potentially subvert a similar fraction of arXiv's operating costs. It would be unreasonable to expect SCOAP3 or HEP labs to cover the entire cost of arXiv.

Another question we often received during our initial assessment process was the relationship between arXiv and other related databases such as Inspire and the Stanford Physics Information Retrieval System (SPIRES).[2] arXiv has long enjoyed close ties with the SPIRES and Inspire initiatives and project partners at the European Organization for Nuclear Research (CERN), SLAC, Deutsches Elektronen-Synchrotron (DESY), and Fermilab. We expect this collaboration to result in improved services for the HEP community, and we continue to investigate ways to share tools and software. However, the scope of INSPIRE is narrower than arXiv and we do not expect to find significant savings in operating costs. We continue to investigate additional partnerships that will enable us to improve the discovery and interoperability features of arXiv. We also perceive such partnerships essential for bringing cost efficiencies.

3.2. Short-Term Business Plan

Currently, we are implementing an interim business arrangement (2010-2012) that aims to generate funds through recurring subsidies from the libraries at academic institutions, research centers, government laboratories, and other organizations that are the heaviest users of arXiv. The model entails a tiered structure of annual support requests similar to many other open-access funding models. The tier-based support structure is based on the previous calendar year's download activity and is applied equally to academic institutions, research centers, government labs, and other organizations. The 3-tiered institutional support model suggests institutional contributions within the range of $4,000 and $2,300 per year. We seek support from institutions representing the most active users of arXiv, in both the United States and other countries. Cornell University Library will continue to provide 15% of arXiv's operating budget, an amount many times higher than the support we will request from other heavy user institutions.

The calendar year 2010 budget for arXiv is $400,000, which includes costs for personnel and operating expenses [4]. The operational expenses include server ware, backups, storage, and preservation services. Staff salaries account for nearly 80% of total annual expenses. Running the repository involves 4.66 FTE staff, including user support, programming, system administration, and management. With over 60,000 new submissions per year one may think of this as an effective cost of approximately $7 per submission. Alternatively, with over 30,000,000 full-text downloads

[1] http://scoap3.org/

[2] Inspire is a HEP information system that aims to integrate existing databases and repositories to host the entire corpus of the HEP literature worldwide. Run by the Stanford Linear Accelerator Center (SLAC), the SPIRES is a database of particle physics literature.

per year this is an effective cost of approximately1.4 cents per download.

We have no plans to impose article processing charges or submission fees. Barrier-free submission and use is one of the founding principles of arXiv. We have considered requesting donations at time of submission but have concluded that such fundraising would incur greater overhead than the institutional support model, and would not engage our peer institutions. We also want to ensure broad international contributions to the repository without financial expectations from the authors. We are committed to maintaining arXiv as an open-access resource that anyone may use to download and read articles as well as allowing submissions free so that all appropriate articles can be accepted.

4. TECHNICAL PLANS FOR ENDURING ACCESS AND ADVANCING ARXIV

4.1. Scalable and Expandable Architecture for Sustainability

The arXiv software has been developed in-house over many years and this has both benefits and burdens associated with it. To keep arXiv sustainable, it is important to re-engineer the software to layer arXiv-specific functionality over generic repository software. Creating a generalized architecture will facilitate efficient technology management processes and allow the implementation of digital preservation procedures and policies.

The arXiv software was developed in-house at the Los Alamos National Laboratory and Cornell over the past eighteen years. The software has evolved and predated most other repository systems. It is predominantly written in Perl with components that use Java, PHP and Python. Metadata and user information is stored in a MySQL database and Lucene is used to provide the search service. The three server machines that provide the main arXiv.org site are supported by Cornell's central IT organization with 24x7 support. Mirror sites are locally supported and receive updates daily.

While the underlying technology has been periodically updated, the system requires significant internal re-engineering to support an evolving technological landscape, increased growth and use, and to ensure the sustainability of the service. The arXiv repository architecture includes some elements that are specialized to the user community, for instance the TeX processing system and the optimized administrative workflow in support of submissions and ingest. Other repository features are more generic and are good candidates for replacement with standard components in order to reduce costs and free developers for the development of new features and services.

We are in the process of surveying and assessing repository technologies. For example, one of the options is adopting a very standard platform such as Fedora for underlying repository functionality. Another strategy is implementing a community-based archival solution such as the Invenio system,[3] which is common within the physics community (developed at CERN). Such a system will lend itself for building features that target e-print archives and also will support the development of shared tools and web services that factor in disciplinary scholarly communication patterns and values.

The second element in our future technical plan is digital preservation of arXiv content in order to support the long-term maintenance of bitstreams and ensure that digital objects are usable (intact and readable), retaining all quantities of authenticity, accuracy, and functionality deemed to be essential when articles (and other associated materials) were ingested. Formats accepted by arXiv have been selected based on their archival value (TeX/LaTeX, PDF, HTML, OOXML) and the ability to process all source files is actively monitored. The underlying bits are protected by standard backup procedures at the Cornell campus and off-site backup facilities in New York City provide geographic redundancy. The complete content is replicated at our mirror sites around the world and additional managed tape backups are taken at Los Alamos National Laboratory.

The Cornell University Library is developing an archival repository (to be operational in May 2011) that will support preservation of critical content from institutional resources including arXiv. All arXiv documents, both in source and processed form, will be stored in this repository and there will be ongoing incremental ingest of new material. We expect that the preservation costs for arXiv will be borne by the Cornell University Library leveraging the archival infrastructure developed for the library system. As an interim solution and also a secondary archival strategy, we are also assessing community governed archives such as CLOCKSS and TRAC-certified services such as PORTICO.

4.2. Sustainability through Innovation

Keeping open access academic resources such as arXiv sustainable involves not only covering the operational costs but also continuing to enhance their value based on the needs of the user community. arXiv's success has relied upon a highly efficient use of both author and administrative effort, and has served its large and ever-growing user base with only a fixed-size skeletal staff. In this respect, it long anticipated many of the current "Web 2.0" and social networking trends, providing a framework in which a community of users can deposit, share and annotate content. It also helped initiate an open access to scholarly literature movement and continues to play a leading role in such endeavors.

[3] CDS Invenio digital library system is a suite of applications that provides the framework and tools for building and managing e-print servers. The software is free and can be licensed under the GNU General Public Licence (GPL).

A critical aspect of the arXiv sustainability plan is enabling interoperability and creating efficiencies among repositories with related and complementary content to reduce duplicate efforts. Organizations with institutional repositories are usually keen to have them used, and would like to avoid the need for authors to make multiple deposits. SWORD (Simple Web-service Offering Repository Deposit) aims to lower the barriers in contributing content in multiple repositories [5]. arXiv implemented the SWORD protocol for automated deposit over a year ago. This protocol enables both multiple deposits from a single tool and deposit from another repository. However, it has yet been used to address the 'multiple deposit problem' while it has been successfully used by journals and conference systems depositing in arXiv.

Digital data and associated multimedia information such as images and audio/video are becoming an integral part of scientific publications. To maintain its innovation role in scholarly communication, it is essential for arXiv to develop features in support of the deposit and archiving of supplementary information objects that are associated with a given paper. Also critical will be to factor in such multimedia content in the development of our preservation plans.

The Cornell University Library frequently receives requests to extend arXiv to include other subject areas. Due to limited resources, we have adopted a measured approach to expansion because there is significant organizational and administrative effort required both to create and to maintain new subject areas. Adding a new subject area involves exploring the user-base and use characteristics pertaining to the subject area, establishing the necessary advisory committees, and recruiting moderators. Although arXiv.org is the central portal for scientific communication in some disciplines, it is neither feasible nor necessarily desirable to play that role in all disciplines. However, arXiv can provide a model for other communities through improved service to its existing dedicated user communities, and act as an essential component of a global networked scholarly communication system. We anticipate that system will become increasingly broad in its subject area coverage, and increasingly diverse in its component databases, repositories, and other online tools and services.

5. PLANS FOR DEVELOPING A LONG-TERM SUSTAINABILITY STRATEGY

We realize that our business model needs to be responsive to the shifting ecology of scholarly publishing. Our current sustainability model represents a short-term strategy for the next three years. We are collaborating with the heaviest user institutions in the US and abroad in our effort to reposition arXiv as a vested online scholarly resource, an asset with shared benefits and accountability. We are very pleased that so many institutions have already stepped forward to share the cost of arXiv. As of June 2010, 79 institutions have pledged their support totaling to $283,000 in contributions. Complementing this short-term business planning efforts, we formed an international advisory group, which will provide an essential consultative role in developing diverse sustainability strategies for this critical international resource.

Over the next few years we will develop a long-term business plan that provides a strategic framework to protect and increase the value of arXiv for those who use it. Ideally this will comprise a blend of ongoing underwriting from Cornell University Library and support from the academic library community and research centers. It might also include support from scholarly societies, an endowment, or funding agencies such as the NSF. We will strengthen existing collaborations (e.g. with the INSPIRE project of CERN, SLAC, DESY and Fermilab) and develop additional partnerships that allow arXiv to provide better services or to share the support burden. Advice from the sustainability advisory group and other supporting institutions will be used in developing this long-term business plan. Also critical for our long-term strategy is developing an approach for reliable and committed stewardship in order to sustain the technical operation and innovative track record of this highly valued repository.

6. CONCLUDING REMARKS

The arXiv case study presented in this paper illustrates the need to approach digital preservation of repositories holistically by taking into consideration a range of lifecycle and usability issues as well as factoring in the changing patterns and modes of scholarly communication. As we collectively address the creation and management of community-based infrastructures, we need to factor in financial needs, usability factors, innovation in discovery and access, and enduring access. arXiv complements, rather than competes with, the commercial and scholarly society journal publishing market. A critical question for the repository and preservation community to address is the versioning of scholarly articles, from initial submission to pre-print archives to their final publishing in formal scholarly journals.

One of the goals of our business planning initiative is to provide a case study that can be used by other institutions with similar repository responsibilities. As support for open access publishing increases and the reliance of users on free resources grows, it is inevitable that educational and cultural institutions will need to collaborate in experimenting with different funding strategies. What is essential is for organizations with such undertakings to share their experiences and lessons learnt with the broader community in order to collectively enhance our understanding of issues and pros and cons of potential strategies. To this end,

Cornell University Library is committed to continue discussing the sustainability planning process and outcomes with our colleague through different forums and channels.

7. REFERENCES

[1] arXiv Business Model White Paper, January 2010. http://arxiv.org/help/support/whitepaper.

[2] Guthrie, K., Griffiths, R., Maron, N. *Sustainability and Revenue Models for Online Academic Resources. An Ithaka Report*. 2008. http://www.ithaka.org/ithaka-s-r/strategy/sustainability-and-revenue-models-for-online-academic-resources

[3] Raym Crow. *Income Models for Open Access: An Overview of Current Practice*, 2009.http://www.arl.org/sparc/publisher/incomemodels/

[4] aXiv 2010 Budget, June 2010, http://arxiv.org/help/support/2010_budget

[5] SWORD (Simple Web-service Offering Repository Deposit) Deposit Lifecycle White Paper. http://www.swordapp.org/.

AUSTRIAN STATE RECORDS MANAGEMENT LIFECYCLE

Berthold Konrath
Austrian State Archives
Nottendorfer Gasse 2
A-1030 Vienna
Austria

Robert Sharpe
Tessella
26 The Quadrant
Abingdon Science Park
Abingdon OX14 3YS
UK

ABSTRACT

The Austrian state is building an integrated "cradle to grave" electronic records management and archive process to ensure that electronic records are managed correctly throughout their lifetime.

This has already included the rollout of records management through federal agencies via the ELAK (Elektronisher Akt) system and the specification of a format for transfer between agencies called EDIAKT. The latter includes transfer to the national archives and thus, in essence, the definition of the format of a valid SIP that can be ingested into an archival system.

The Austrian Federal Chancellery is now funding the provision of such a central archival system plus a general license allowing all Austrian public bodies to benefit from the technology for archiving and preservation. After a competitive tender, Siemens IT Solutions & Systems are providing this system utilising the Safety Deposit Box (SDB) system from Tessella.

This system will ingest the SIPs (in EDIAKT format) into long-term storage and provide comprehensive access, data management, preservation and administration functions. The Österreichische Staatsarchiv (Austrian State Archives) will be the first to use this system by the end of 2010.

This is the basis on which the requirements for maintaining electronic records, which will be the sources of historical research in the future, are being created. This will preserve the historical heritage of Austria for generations to come.

1. INTRODUCTION

The impact of modern technology (computers, mobile phones, the Internet etc.) has been felt throughout our daily lives for some time now. As part of this trend there has been a huge impact on government departments and consequentially on government records. This has spawned new phrases such as "e-government" which in some ways has become a synonym for a modern state. However, this exciting trend also throws up challenges and it forces records mangers and archivists to have new processes.

For years, and increasingly so in the recent past, modern administrations have used IT-instruments in the fulfilment of their statutory tasks. As early as in 1985, the Austrian Archiv der Republik (the part of the Austrian State Archives [1] responsible for records post 1918) introduced an electronic file administration system. In 2003 the Austrian State Archives were among the first Austrian federal services to change over to the exclusive use of electronic files.

In Austria, the long-term storage of electronic data from both electronic file administration systems and other systems as well as the acceptance of "traditional hard copies" are responsibilities of the Archiv der Republik. In this context, experts from the Archiv der Republik have, from the very start, been involved in the introduction of the "paperless office" (use of federal electronic files), the management of electronic files (Document Lifecycle Management) and the creation of an electronic interface (EDIAKT II) between the different electronic filing systems of the Austrian federal administration.

More recently, a feasibility study was undertaken in 2006-7 to define requirements, possible solutions, conventions and categories for a digital long-term archive. One of the key requirements was that the system needed to be compatible with key international standards especially OAIS. This led to a clear need for cooperation between the Austrian State Archives and the Austrian Federal Chancellery in order to procure the system called Digitale Langzeitarchivierung im Bund (DigLAimBund) based on these requirements.

This led to a public tender which was won by Siemens IT Solutions & Services together with their software partner Tessella utilising the Safety Deposit Box (SDB) system which, in conjunction with appropriate hardware and other systems, constitutes an OAIS-complaint solution. This system will be used by the Österreichische Staatsarchiv (Austrian State Archives) and other agencies in order to ensure the preservation of electronic records for the next generation.

2. FEDERAL RECORDS MANAGEMENT

As experience has too often shown, paper files can be lost, misplaced, incorrectly filed, or land in a back corner of the archives. Hence, one of the most important developments of eGovernment for the Government is the electronic record system, called ELAK. It enables seamless communication between public authorities and other governmental services.

In 2001, the ELAK ("Elektronischer Akt) system for records management was launched department-wide in the Austrian Ministry for Foreign Affairs and the Federal Chancellery. Since then, ELAK has been rolled out nation-wide and is also being introduced step by step in provincial governments.

The advantages of electronic record processing are obvious. ELAK substantially reduces the amount of time required for processing applications since documents no longer need to be sent back and forth between ministries and public authorities. Instead, they can be processed conveniently online. Processes are standardized and can run parallel to one another. Enquiries can be carried out directly from the desktop and the process workflow is completely transparent. With practically just a push of a button you can find out at any time of day how far the file has been processed. Furthermore, there are never any problems due to changes in the format of the file (printed copies, scans) because ELAK is based on a standardized system with uniform user interfaces. The days of traditional paper-oriented file processing are numbered. In the meantime, paper-oriented file processing is being replaced by automated business processes.

In their function as a document and workflow management system for the electronic implementation of internal work processes, electronic file systems become a kind of data hub in which different applications and data sources can be integrated so that changes in media format can be avoided. In the electronic record system of the federal administration, the most important interfaces and systems for public administration are:

- Form server: This interface displays forms in graphical user interfaces, making it the most important interface from the citizen's point of view. Application forms that are submitted over a Web form can be processed directly in the ELAK system due to their standardized data structure and XML syntax.
- Electronic delivery: In order to transmit information, notices, and documents to the intended person, the piece of correspondence must be delivered via a delivery service using the methods described.
- Interfaces to other applications: information is often needed from citizens during procedures which they are not able to supply, either because it would require too much effort, or because it may not be possible for the citizen to do so. Instead of citizens having to chase their data around, the data should be able to be accessed by the ELAK system in an automated manner from public administration applications such as registers, SAP systems or directory services. Communication occurs over defined interfaces that support the standardized exchange of data.

The Austrian eGovernment strategy requires active participation in creating interfaces which are standardized across public authorities and drafting specifications that are effective nationwide as part of the cooperation between the Federal Government, the provinces and municipalities. The results from the work groups are based wherever possible on international norms and standards, or use them as a model. The typical eGovernment components that are needed in administrative and back-office processes join together to form a big picture. Along with individual applications, the big picture includes modules for online applications and components of the citizen card concept. The protocols used in the communication architecture function figuratively as the mortar that holds the building blocks together.

3. DEFINING SIPS USING EDIAKT

The ability of government agencies to use and manage electronic records is just a start. It is also necessary for these agencies to exchange information with each other. Although all such agencies have record management systems that work with electronic records, records of business processes, and sub- processes including documents, the objects were specific to the manufacturer of the software and not built according to a uniform standard. To this end, EDIAKT [4] was developed as a format for standardising communication between different public institutions (authorities, courts of law, businesses).

In the course of further development of the EDIAKT system and due to increased distribution of the ELAK system, the standard was updated to its current format, EDIAKT II.

In this standard, data is packaged as EDIAKT objects, which are comprised of:

- Meta-data that describes a record, business (sub-) process, or document
- Process data for process instances and activities in accordance with the XPDL standard of the Workflow Management Coalition
- Content of the record, business (sub-)process, or document
- Procedure-specific data that may be attached to an object.

To satisfy the different requirements of institutions using ELAK, EDIAKT implemented a hierarchical structure with four layers. At the bottom is the

document, which contains the file in its original format. If the file is not saved in a standard format, a document with a standard format must be attached. One or more documents are encapsulated in a record of a business sub-process. It represents the smallest object in EDIAKT II. This business sub-process may further be aggregated along with other sub-process in a higher-level business process record. Authorities that do not have their own ELAK system can still read EDIAKT packages using the free EDIAKT Viewer. The current version can be used to:

- Display all meta-data including process data,
- Show embedded documents,
- Verify digital signatures.

EDIAKT II is used more than just as an interface between different electronic record systems. It can also be used for internal data exchange between special applications and archive systems. EDIAKT II, together with the EDIAKT Viewer and EDIAKT Creator, and supplemented by the standard document format PDF/A, establishes the basis for the long-term archiving of records. In the future, this format could play an increasingly central role for the submission of original records that is required for different courts of jurisdiction.

4. DigLAimBUND OVERVIEW

Once government records have reached the end of their operational life within the original creating agency (and other related or successor agencies) they need to be assessed to see if they need to be retained for a longer time period. A proportion of those retained will indeed eventually be selected for permanent retention at the Österreichische Staatsarchiv (Austrian State Archives) and will thus be transferred to them.

This requires the Austrian State Archives to have a system that is capable of ingesting, storing, managing, preserving and providing access to these records. This is the role of the Digitale Langzeitarchivierung im Bund (DigLAimBund).

DigLAimBund is based on the pre-existing SDB4 system. This system is a Java web-based server application running with a relational database (in this case Oracle) behind it. When combined with operational hardware, a physical storage system and a system for authorisation and authentication, it offers all the functions required of an OAIS system. In the following sections for the rest of the paper each of the functions of each of the functional entities in OAIS are discussed in turn illustrating how DigLAimBund complies with this.

One of its key features is its "Active Preservation" module that allows automated, verifiable digital preservation to occur controlled by a Technical Registry. The Technical Registry is another Java web-based server application with a relational database behind it. It is based on the Planets Core Registry [8] that is itself based upon the UK National Archive's PRONOM system [4,5]. The Technical Registry contains not just factual information (e.g., a list of formats, known software, migration pathways etc.) but also policy information (e.g., which formats or combination of formats and properties are considered to make a file obsolete and thus in need of preservation action, how to measure these properties, which preservation action to perform in which circumstances etc.). This policy is machine-readable which is the key to allowing digital preservation to be automated.

Another feature is that it contains a built-in workflow system, which is based on Drools Flow (an open-source development). This allows configurable workflows to be created with comparative ease for each OAIS processes. Each workflow consists of a series of workflow steps. These steps can be automated steps or can involve user input (via a web form). Each step is self-recording so that an audit trail of actions is created.

5. DigLAimBUND INGEST

The first ingest function defined in OAIS is the ability to receive a submission. Clearly this requires a government agency to transfer a SIP in the EDIAKT II format to the archive.

5.1. Receive Submission

Once the physical transfer has taken place, it needs to pass into the boundary of the DigLAimBUND system. The first step to be performed is to transform the EDIAKT package into a format that is understood by the archiving system. This allows the SIPs to then be converted to AIPs in such as way as to utilise standard ingest, storage, access, data management and preservation workflow steps already present in SDB. This restricts the amount of development needed within the project to those aspects needed for genuinely local configuration or enhancements.

In the case of SDB, it utilises a metadata schema called XIP (that covers SIPs, AIPs and DIPs). This schema defines the structural metadata needed to link records to files in given manifestations. This is especially important for EDIAKT II since it is normal to receive both an original and a normalised form of each record (e.g., Word and PDF documents). In addition, the XIP schema defines technical metadata. In particular, it has been specifically designed to allow the automation of digital preservation (see below). However, XIP does not proscribe descriptive metadata, instead allowing descriptions to be described using any appropriate metadata schema (e.g., EAD), which can be embedded inside XIP.

5.2. Perform Quality Assurance

Next the "perform quality assurance" steps required by OAIS are performed. This includes virus checking, verifying compliance with schemas, fixity value checks and a check that every entity's identifier is unique.

5.3. Generate AIP

The next step is to generate the AIP. As described above, XIP allows this to happen through gradual refinement but one of the key steps involved is characterisation.

In SDB characterisation is itself a multi-stage process involving both technical and conceptual characterisation. It is a fully automated process that is designed to allow future preservation processes to also be fully automated.

5.3.1. Technical Characterisation

Technical characterisation involves discovering properties of the actual files with the aim to discover those properties that might determine whether the file is in an obsolete technology and thus in need of some form of preservation action:

- First of all it attempts to identify the format of each file using DROID [3]. Importantly, this links the file to a format identifier, which can be used to automate preservation policy decisions based on information stored in the Technical Registry.
- Format validation then takes place. The initial format identification determines the best validation tool to run (e.g., Jhove [6] is run for JPEG files). (The policy decision of which tool to use for which initial format identification is stored in a machine-readable way in the Technical Registry thus enabling this process to be automated). This can lead to the format identification result being updated. For example, DROID is currently unable to distinguish TIFF3, TIFF4, TIFF5 and TIFF6. However, Jhove is capable of validating each of these formats and will thus be able to reject three of the four initial identifications.
- Each file then has key properties extracted by means of a tool. The tool used and the properties extracted are again format dependent and are determined by the policy in the Technical Registry. Again, importantly, each property is linked to a Registry identifier so that any policies associated with that property can be automatically applied.
- Where possible, embedded objects are extracted from each file and these are characterised in turn. This is important because the embedded object may be obsolete even if the container file is not. (Once again the tool to use to perform this extraction is based on the format of the container file and is based on machine-readable policy stored in the Registry).

5.3.2. Conceptual Characterisation

Conceptual characterisation determines the conceptual units called "components" that need to be preserved. These are not necessarily equivalent to files since, being conceptual, they are not technology-dependent. For example, one component might be a "web page" which, in current technology in 2010, is likely to consist of many files (HTML, CSS, GIF etc.) that combine to produce a conceptual entity that needs to be preserved. However, there is no guarantee that the physical structure will be identical in future generations of technology.

Once these have been identified the technology-independent properties of these components should be measured. These form the "significant properties" of the component that should be invariant in a good migration. A record will be well preserved if all its components, all their properties and all the relationships between these components are preserved.

In practice, of course, the conceptual properties need to be measured in the technology present in the SIP so component properties are closely linked to the technical properties measured for individual files (or an aggregation thereof). However, the distinction between them is important even if there is often a one-to-one correspondence between a file and a component in current technology: file properties are technology-dependent and thus needn't necessarily be preserved while component properties are technology-independent and thus should be preserved. This will be discussed more in the preservation section below.

5.3.3. Quality Assurance Revisited

In practice, there is an overlap between the steps involved in generating the AIP and performing quality assurance. For example, quality assurance restrictions on permitted formats or allowed properties of files (e.g., preventing encrypted PDFs from being ingested) can only be applied after technical characterisation has taken place. Also, some of the steps listed in section 5.2 (e.g., virus checking) lead to metadata (such as information on the virus checker used) being added to the AIP.

5.4. Generate Descriptive Information

Part of the OAIS ingest process requires the system to ensure that all the systems that need to hold descriptive information are synchronised. The Austrian State Archives maintain a catalogue that contains descriptive information about all of their holdings, whether this is on traditional media or electronic. Hence, it is necessary for DigLAimBund to be able to produce a snapshot of the descriptive information needed by a catalogue

system and making it available to that system. This is done using OAI-PMH.

5.5. Coordinate Updates

The last step of ingest defined in OAIS it to send the AIP to be stored in the combination of the relational database and the bulk-file storage system. This is described in more detail in the next two sections.

6. DigLAimBUND STORAGE

6.1. Receive and Provide Data

The bulk-storage system used in DigLAimBund is EMC Centera. SDB interfaces to such a bulk-storage system through a series of APIs that isolate changes in the storage system and changes in the repository software. There are interfaces to allow content to be stored and retrieved in a variety of ways (e.g., with or without a metadata snapshot, with content files stored independently or within a package, whether to sign a package or not etc.). Each of these decisions will be discussed in turn.

Metadata is stored in the database so, if this is properly backed up, storing a metadata snapshot may seem to be unnecessary. However, adding such a snapshot means that in the event of a non-recoverable database failure the storage system contains enough information to restore a record to a known state. On the other hand, it should be noted that the database contains the latest set of information about the record (e.g., information on access events) so some information will be lost if the database is lost. Of course, this information could be stored in the bulk storage system as well if the snapshots are refreshed at regular intervals but this would place quite a burden on the storage system. DigLAimBund has opted for a reasonable middle ground and does add a metadata snapshot to the storage system but will only update it if a preservation action occurs: not in the event of an access event or a descriptive metadata update. Note that any descriptive metadata updates will also be stored in the catalogue system so they are backed up independently anyway. This means that the historical information of who accessed the record when would be lost in the event of a non-recoverable database failure but this seems to be a reasonable compromise.

DigLAimBund also stores AIPs as packages (one AIP per SIP received). This is partly a policy decision and partly a consequence of the storage system which, if used to store a lot of small files (as might be the case when storing a web site), will waste a lot of expensive storage capacity.

Finally the packages are signed with a XadES signature, which is used to further guarantee the authenticity of the package.

6.2. Managing Storage

OAIS requires the system to manage the storage hierarchy, replace media as required and provide disaster recovery. All of these features are provided through the standard features of EMC Centera.

6.3. Error Checking

EMC Centera provides built-in features that check every file against its fixity values in order to pick up any corruption. In addition, SDB provides an on-going integrity checking function (based on a least recently checked algorithm) that does the same across as many storage adaptors as the system has (this allows for, for example, a second copy to be stored in different storage technology). A further advantage of this duplication is that the SDB check also provides a means for checking that the list of files held in the metadata database and those actually stored are identical (and that the fixity values stored in both sub-systems are also the same).

7. DigLAimBUND DATA MANAGEMENT

7.1. Receive Database Updates

Database updates (whether ingest requests or update requests for a variety of reasons described below) are received by the SDB database and processed by storing entities from the data model into appropriate database tables with all the information held in an XML fragment and some information denormalised into standard relational database fields where fast access or querying capability is required.

7.1.1. Post-Ingest updates

SDB provides the ability for descriptive metadata to be enhanced. However, it is also possible for this metadata to be updated in the catalogue system. Hence, exchange of information between the systems (via OAI-PMH) is very important.

In addition, DigLAimBund supports a few specific scenarios:

- Allow records to be moved to a new collection.
- Allow records to be appraised after ingest and then, if necessary, to be exported (in EDIAKT form) for ingest to another system or to be deleted altogether.
- Allow records to be deleted as a result of a court order.

Appraisal and deletion actions occur via a “four eyes” principle (i.e. the workflow requires a supervisor to approve an initial assessment) while deletion via a court order (a very rare event) will occur via a careful operating procedure. In order to support this SDB also includes the ability to “soft delete” (i.e. to immediately prevent the record from being visible to ordinary users while the full workflow is enacted).

Finally, updates can occur as a result of preservation actions (see below) or re-characterisation (re-running characterisation to take advantages of better tools).

7.2. Perform Queries

All database accesses utilise Hibernate [5] so that the system is not dependent on any particular database engine technology (although Oracle 11 is used). This means that all queries onto the system work using HQL rather than SQL. All queries needed for operation of the system in normal circumstances are already built-in to SDB.

Of course for efficient querying it is necessary to use appropriate indexing. Hence, in addition to standard relational database indexes, SDB uses the Solr [10] search engine to index the descriptive metadata held in XML fragments and to perform full text indexing of the (text-based) content files.

7.3. Generate Report

Reporting can be made in two ways in DigLAimBund: internal SDB reporting using the open-source Jasper Reports tool (which requires some programming ability but allows reports to be embedded within the application) and an external reporting tool using the Pentaho reporting tool (which allows simple reports to be created in a less technical way). In either case full access to the entities held in the database is provided including the audit trail and the workflow history so that the full provenance of every entity can be reported upon.

7.4. Administer Database

This uses standard database tools provided by Oracle.

8. DigLAimBUND ACCESS

8.1. Coordinate Access Activities

8.1.1. Query Requests

DigLAimBund provides the ability for users to:

- Browse a tectonic to find a record of interest
- Search for records by simple search (across all information held) and by advanced search (i.e. by choosing the appropriate fields). This includes the ability to search within the full text of documents. Each search identifies records that match the criteria order by relevance and (where full text searching has occurred) identifies the documents within that record responsible for the hit.

Access is only provided to records that are within the rights of the individual user to view. Once a user has found a record they can view all the metadata known about it (including its place in the tectonic and descriptive metadata). They can also see the list of files held together with (for common formats) a snapshot of the file.

For archival staff all the information held in the metadata store is available including:

- The list of possible manifestations available for download
- For each manifestation, the list of files and the list of components (identified in conceptual characterisation) that constitute it.
- For each file, all the technical metadata held
- The full audit trail for each entity held.

8.1.2. Orders

Authorised users can order content in two ways: an ad-hoc order (immediate download or rendering of a single selected file) or an event-based order for a record.

8.2. Generate DIP and Deliver Response

When an order is received, the appropriate content is retrieved from storage and a DIP is generated. This requires the appropriate files to be retrieved from storage, their integrity checked and then to package them up into a package (e.g., a ZIP file). This is then delivered to the end customer. For event-based orders, this can take place in a number of ways (e.g., via a download, by e-mail or placing the content in an pre-assigned location and informing the end user).

9. DigLAimBUND PRESERVATION

9.1. Preservation Planning in OAIS

Most of the preservation activities required in OAIS are to do with planning rather than performing preservation and are mainly activities requiring human judgement.

These are, of course, very important activities. However, one of SDB's (and thus DigLAimBund)'s main features is "Active Preservation" (an automated way of performing preservation). This is explained in this section.

9.2. Policy

The Technical Registry contains information about, amongst other things, formats (and format technical properties) and migration pathways. This allows policy to be set about what makes a file obsolete and what to do to migrate files to a new format. This can be either an absolute measure (e.g., a statement that any file in a given format is considered obsolete) or a risk-based measure (e.g., a series of a criteria that contribute towards risk and if, when taken together, pass a threshold, would make a file be considered obsolete). The Registry also allows policies to be set for different reasons (e.g., obsolescence of the preservation copy could follow a different policy than obsolescence of the presentation copy).

The Registry can be used in two ways: either by allowing policy approval so that the official policy governing one particular scenario is clear or by allowing

manual intervention in the otherwise automated preservation workflows to pick the policy appropriate to the particular scenario. In reality a combination of these approaches is in use as best practice in this area is still in development.

9.3. Determining Files and Records at Risk

The policy criteria that determine obsolescence are stored in a machine-readable way which means that they can be automatically compared to the technical characteristics derived during ingest (or a subsequent re-characterisation) in order to determine which files are in need of action. It is then possible to work out which manifestations of which records within the repository are in need of some form of preservation action. In order to identify which manifestations are relevant, each manifestation of a record is typed (e.g., "preservation" or "presentation").

This process can take place at any time so, in order to prevent repeated migrations, each manifestation is also assigned an active flag. This is set to ensure that there is only one active manifestation of each type allowed at any one point in time and only active manifestations of the type that corresponds to the migration reason (and thus to the stored policy) are considered for migration.

9.4. Extending to Linked Records

Having established which record manifestations need attention, the system then extends the migration to include all other records within the branch of the tectonic. This is since, for example, a parent record of a record in need of attention needs to be deliverable in full in the new manifestation. Hence, it is essential that the system checks that the new manifestation of the parent (which will include the files of the child record) is coherent. This may or may not lead to any additional file migrations but it will lead to additional verification checks if there are links between the records. As an example of this, the parent record could be a web site and the child record could be a report held within that web site. If the report were migrated from, say, Word to PDF leading to a change in file name extension, the html page of web site would need to be slightly altered in the new manifestation so that it links to the new file.

9.5. Migration

Having determined the extent of migration needed, the system then determines all of the components of the records discovered during conceptual characterisation (described above). Some of these components will contain files that need to be migrated (either because of obsolescence or because of the knock-on effects such as the web page described above). These are the atomic units of migration since these units and their appropriate properties and relationships are the things that are preserved during migration even though the physical structure of the files that manifest them may change.

Based on the policy described above, each such component runs through a migration pathway, which determines the migration tool(s) to use, thereby migrating the set of files it contains into a new set of files. The new files produced then run through technical characterisation and the new component manifestation through conceptual characterisation. This latter step identifies the new component manifestation's properties and relationships that should be identical to those in the original (subject to any tolerances permitted in the policy owing to, perhaps, acceptable rounding errors or an expected degradation such as a lower resolution image being created for preservation). All of this information is held in XML thereby again allowing this process to occur automatically.

Once this process has been completed successfully, the system can aggregate the component manifestations into record manifestations and ingest these into the repository. The system also ensures that the superseded manifestation is turned inactive so it is not migrated again.

9.6. Alternative Approaches

As described above, SDB currently supports the case of "just in case" migration. However, "on demand" migration would be possible by adding an appropriate workflow. This could simply be access workflow (i.e. create the migrated copy and provide it to the end user) or could be a full preservation workflow using "Active Preservation" if the intention was to ingest the ne manifestation in addition to providing immediate access).

In addition, trial emulation functionality is currently being added to SDB as part of the EU-funded KEEP project [7].

10. DigLAimBUND ADMINISTRATION

DigLAimBund includes sophisticated administration features to perform the features required by OAIS namely:

- Manage the system configuration. This includes performing standard IT system administration (e.g., monitor backups, database performance etc.).
- Establish standards and procedures by configuring the workflows for ingest, access, storage, data management and preservation. Workflows can be started manually, at regular intervals or in response to monitored events such as the arrival of a SIP in a specific location. If a step requires individual attention the appropriate user is informed via e-mail. Each user when they log-in can see a list of any actions awaiting their input.

It is also possible to report on the progress of workflows or to monitor what happened in workflows that have been completed at any time in the past.

- Control access rights
- Allow archival information updates (e.g., metadata editing, deletion and appraisal as described above)
- Audit information (e.g., to allow users to report on the contents of the archive or an authorised user to view the audit trail of any entity in the system).
- Negotiate submission agreements. Transfer agreements (including restrictions on SIP sizes, allowed formats etc.) can be set-up and automatically verified during ingest

11. CONCLUSION

The Austrian State has been investing heavily in electronic records management and archiving. This has already led to the use of records management within government agencies (via ELAK) and a system for transferring material between agencies (using EDIAKT).

It now also includes an archival system (DigLAimBund) that will be operational in late 2010.

Hence, much work has been done but it is anticipated that further work will be needed especially in the establishment of the best practice that is needed to run the system efficiently. In the interests of developing and sharing this, Tessella and institutions that utilise the SDB system have formed an SDB Users Group that has already met on four occasions to participate in this exchange of hands-on information.

12. REFERENCES

[1] Austrian State Archives: http://www.oesta.gv.at

[2] Brown, A. "Automating preservation: New developments in the PRONOM service", *RLG DigiNews*, 9(2), 2005.

[3] DROID home page: http://droid.sourceforge.net

[4] EDIAKTViewer/Creator, http://www.ag.bka.gv.at/index.php/EDIAKT-Viewer

[5] Hibernate home page: http://www.hibernate.org/

[6] Jhove home page: http://hul.harvard.edu/jhove

[7] KEEP project home page: http://www.keep-project.eu/

[8] Planets project home page: http://www.planets-project.eu

[9] PRONOM home page: http://www.nationalarchives.gov.uk/pronom

[10] Solr home page: http://lucene.apache.org/solr/features.html

Session 6 (Panel): How Green is Digital Preservation?

PANEL DISCUSSION: HOW GREEN IS DIGITAL PRESERVATION?

Neil Grindley
JISC
London, UK

Kris Carpenter Negulescu
Internet Archive
San Francisco, USA

William Kilbride
Digital Preservation Coalition
York, UK

Diane Macdonald
University of Strathclyde
Strathclyde, UK

David Rosenthal
LOCKSS
California, USA

ABSTRACT

Digital preservation practitioners, for the most part, regard themselves as the custodians of our digital legacy, identifying with, and in some cases updating library and archival roles to ensure the safe long-term stewardship of digital assets. Outside of the digital preservation community, it is quite possible (or even probable) that preservation is construed as a mindset where the principal goal is to devise ways of keeping as much digital material as possible in perpetuity. It is only a short step from this assumption to arrive at the conclusion that the whole preservation enterprise is not only environmentally reckless in its ever-increasing demand for server and storage space, but more fundamentally chaotic in its aspiration to defy the capacity of digital librarians, archivists and data managers to keep the social, cultural, scientific and scholarly record well ordered and categorical.

1. INTRODUCTION

There are a number of ways that the objectives of digital preservation may be interpreted, and these match the myriad motivations of those who preserve materials, both for themselves, or more often, on behalf of various designated communities who have decided to entrust their long-term investment in digital assets to expert practitioners. In most cases, the motivations to preserve are transparent and commendable, and are borne out of a positive desire to ensure that subsequent generations have the opportunity to creatively engage with our digital legacy.

This panel is not an attempt, therefore, to brand any parties or processes as being 'anti-green'. Preservation is not primarily designed to address the environmental agenda; its principle purpose is to ensure continuity of memory. However, environmental questions can provoke emotive responses and there is a risk that if the preservation community does not rehearse effective responses to the potential charge of being uninterested in environmental issues, due to its apparent objective to keep ever larger quantities of digital material for the foreseeable future (demanding ultimately unquantifiable amounts of electrical power), the preservation and environmentalist communities are set to collide.

2. INFORMATION LIFECYCLE MANAGEMENT

One issue that the panel might have to address is where the points of engagement between preservation and green issues actually are. Broadening the scope of preservation concerns to encompass aspects of information management brings environmental issues more clearly to the fore. A recent JISC-funded study, Greening Info Management (http://bit.ly/bz2mJL) looks at the potential value of using information lifecycle management techniques to reduce the overall quantity of data requiring management/preservation, thereby cutting personal and organizational energy consumption. Some of the key points of this report will feature in the discussion.

3. THE PANEL

The panel brings together a variety of opinion and expertise from the US and UK.

"On the whole, we probably don't need to question the motivations to preserve, but it may be timely to think

about our objectives. We need to ensure that institutional information policies and strategies are fit for purpose in the face of climate-change and other environmental imperatives." (Neil Grindley)

"It has been the experience of the Internet Archive, that if an institution does NOT address issues of power consumption, etc. there will be a much more limited volume of digital preservation and access that the institution will be able to support over time. The primary operational costs beyond labor costs are usually what you pay for the resources you consume." (Kris Carpenter)

"Digital preservation is used to the idea of managing long term risks so we should be predisposed to thinking about long term environmental risks. As well as being inherently sensible it will become more important in our attempts to influence policy. At the very least we need to be certain that our chosen solutions do not inadvertently become part of the problem." (William Kilbride)

"Holistic positions [are needed] which address the Green agenda and ensure effective stewardship of resources. There appears to be a gap between the well-established understanding of information management per se, and its potential importance for furthering the efficiency of energy usage within the HE and FE sectors." (Diane MacDonald)

"Current hard disk storage is becoming the major consumer of power in data centers. This is both a problem for digital preservation, and an opportunity. Technological changes in the pipeline are likely to both slow the decline in hard disk costs and increase the competitiveness of alternative storage technologies." (David Rosenthal)

4. BIOGRAPHIES

Neil Grindley, JISC (the Joint Information Systems Committee), 5 Lancaster Place, London, U.K.
Neil Grindley is the Digital Preservation and Records Management Programme Manager with JISC and is responsible for a number of projects, studies, and other initiatives that raise awareness and increase the capacity of relevant communities to engage with digital preservation as part of a life-cycle management approach to the creation and exploitation of digital resources.

Kris Carpenter Negulescu, Director, Web Group, Internet Archive, San Francisco, U.S.A.
Kris leads a team responsible for: cultivating IA's web collections and providing access to researchers and the general public; developing the Heritrix open source web crawler and Wayback machine as well as other tools used to search, mine and replay archived web content; providing expertise and services in web archiving, data mining and access to libraries, archives, museums, and memory institutions around the globe.

William Kilbride, Digital Preservation Coalition, York, U.K.
William is the Executive Director of the Digital Preservation Coalition and has a wealth of experience of coordinating cross-domain preservation and archiving initiatives in all sectors of UK activity. His work for the DPC brings him into contact with practitioners and senior decision-makers and positions the DPC as an influential body that helps to shape and refine UK preservation policy initiatives.

Diane Macdonald, University of Strathclyde, Strathclyde, U.K.
Diane is the interim Head of Innovation services at the University of Strathclyde and the lead author of the Greening Information Management Study.

David Rosenthal, LOCKSS, Stanford University, California, U.S.A.
David is Chief Scientist of the LOCKSS digital preservation program at Stanford University Libraries. He is a long-time Silicon Valley engineer who has been researching, writing and speaking on digital preservation for more than a decade

Session 8a: Architecture and Models

FULL PAPER

DIGITAL PRESERVATION FOR ENTERPRISE CONTENT: A GAP-ANALYSIS BETWEEN ECM AND OAIS

Joachim Korb

AIT Austrian Institute of Technology
Safety & Security Department

Stephan Strodl

Vienna University of Technology
Information and Software Engineering Group

ABSTRACT

Today, more and more information is being produced in a digital form. In addition to this so-called born-digital content, material that was produced to exist in an analogue form is now being digitised both for preservation and for easier access. This digital information comes in an ever greater variety of formats, many of which are relatively short-lived. Newer versions of the same software are often unable to render files produced with older versions of that software, let alone files produced with similar software from other vendors. Soft- and hardware environments change constantly and after only a few years can older files often no longer be rendered with up-to-date systems.

While large scientific organisations and memory institutions (museums, libraries and archives) have in recent years invested significant effort and activity towards digital preservation, the commercial world does not currently have the means to preserve their digital information for the long term.

This paper sets out to determine what would be needed to make modern Enterprise Content Management (ECM) Systems ready for long-term preservation of the assets stored within them. For this aim both the general Model of an ECM and the "Reference Model for an Open Archival Information System (OAIS)" have been described, and the special needs of an enterprise system identified.

A special focus lies on the Electronic Records Management (ERM) component of ECMs, which already provides simple preservation functionalities, but lacks those aspects of the OAIS that would make it truly long-term preservation capable. A truly long-term preservation capable ERM would have to add these while retaining capabilities of compliance (the retention or destruction of certain documents in accordance to legal requirements).

1. INTRODUCTION

Over the last decades more and more information has been produced in a digital form. While at first computers may only have acted as 'intelligent typewriters', an increasing part of what used to exist only in an analogue form is now held digitally. Additionally, material that was produced to exist in an analogue form is now being digitised both for preservation and for the purpose of wider access. This digital information comes in an ever greater variety of formats, many of which are relatively short-lived. Oft-cited examples for this are the Microsoft Word format, which has changed continuously over the different versions of the software, and CAD files which are so reliant on the software they were produced with that it is usually impossible to render older files with newer versions of the software, let alone software from a different vendor.

As time progresses soft- and even hardware environments change, so that after only a few years it is often no longer possible to open older files. To understand the problem, two aspects of digital preservation must be considered. First, there is the question of preserving access to the actual bits of digital information; this is usually referred to as bit-stream preservation. Bit-stream preservation includes questions of media integrity as well as the hardware necessary to read the media.

This aspect of the problem is already solved in many ECM solutions, but, while keeping redundant copies of all data may safeguard against the loss of the actual files, it will not guarantee their long-term[1] accessibility. The danger of losing whole collections of data as a result of outdated data formats, software or run-time environments must be countered by developing Digital Preservation Systems – this aspect is known as logical preservation. Logical preservation is the main concern of this gap-analysis.

Large scientific organisations and memory institutions have in recent years invested significant effort into ensuring the long-term availability of their entrusted digital assets. For this, both commercial

[1] [4] defines "long term" as "a period of time long enough for there to be concern about the impacts of changing technologies, including support for new media and data formats […] on the information being held in a repository. This period extends into the indefinite future".

vendors as well as a number of smaller and larger projects[2] have produced systems for storing, managing and accessing those assets. A special focus in all these efforts has been on the compliance with the "Reference Model for an Open Archival Information System (OAIS)" [4]. An ISO-standard since 2003, this reference model not only describes a system long-term preservation but provides a common vocabulary to those concerned with such work.

In the commercial world the situation is quite similar to that described above. Large volumes of born-digital and digitised material are ingested into, managed in and accessed from what is now commonly called Enterprise Content Management (ECM) Systems. The main difference between these and the systems in use in many of the above mentioned institutions is that ECM Systems do not, currently, provide the means to keep the stored information accessible in the long term.

The aim of this paper is to determine what would be needed to make modern ECMs ready for long-term preservation of the assets stored within them. For this aim both types of systems will be compared, the special needs of an enterprise system will be identified and the steps to make a typical ECM OAIS-compliant will be described.

In the enterprise environment there are a number of different terms describing systems used to store digital information. This is mainly due to two facts:

1. These terms were coined to advertise software; and different companies would sell their products under a variety of names to differentiate between theirs and similar products from other companies, but also from earlier versions of their own products that had fewer or different capabilities.
2. Until recently there has been no common model for these systems.

For a long time, the term Content Management was used to describe systems that allowed enterprises to manage document and content flow. Now, however the term Content Management System is most often used to describe software for maintaining, controlling, changing and reassembling the content for internet presentation.

For the purpose of the gap analysis, this paper will follow the Association for Information and Image Management's (AIIM)[3] definitions [2] and use the term Enterprise Content Management System, to describe the strategies, methods and tools used to manage business content. For the description of a long-term preservation archive the Consultative Committee for Space Data Systems' (CCSDS)[4] has produced the OAIS Reference Model. In this analysis, the OAIS terms and definitions will be used when referring to such a system.

[2] E.g. the EU-funded PLANETS project [http://www.planets-project.eu/] or the Austrian RS-DME project [http://www.rs-dme.at/], for which the original version of this analysis was written.
[3] http://www.aiim.org/
[4] http://public.ccsds.org/

The remainder of this paper is organised as follows: Section 2 presents the model of an Enterprise Content Management, followed by the description of the key concepts of the OAIS reference model in Section 3. The results of the gab analysis are presented in Section 4.

2. ENTERPRISE CONTENT MANAGEMENT

The following description of a model for Enterprise Content Management follows the description of AIIM. In 2005, the Association for Information and Image Management, "the leading non-profit organization focused on helping users to understand the challenges associated with managing documents, content, records, and business processes,"[5] set out to find a common name and description to identify the procedures as well as the types of systems that allow for the control of enterprise content.

AIIM describes ECM as "the strategies, methods and tools used to capture, manage, store, preserve, and deliver content and documents related to organizational processes."[2] These tools, methods and strategies are used to manage what is referred to as "the lifecycle" of that content.

For the purpose of this description, the model used here will follow AIIM's descriptions rather than its images. The problem with the images AIIM uses to describe its model is that they are designed to reflect a rather complex concept in a way that makes them perfect for marketing. The complexity of the model is due to the fact that it encompasses strategies, methods and tool, while trying to leave enough room both for the description of different settings and for vendors to emphasise the respective merits of their own software.

2.1. ECM Components

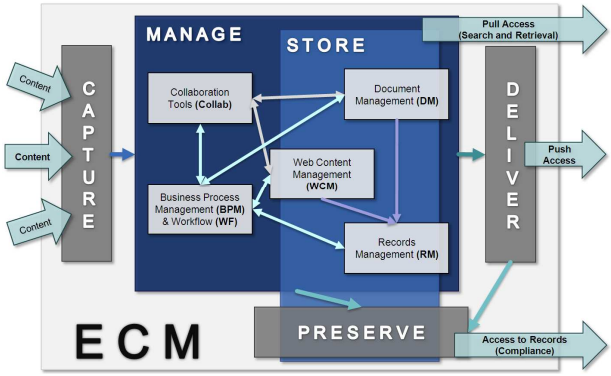


Figure 1. ECM Components Model

The ECM Components model (Figure 1) shows the main components of an ECM, which are:

Capture (Input Management): Capture's function is to ensure that any content within a certain setting is

[5] http://www.aiim.org/AboutAIIM/ECM-ERM-BPM-Association.aspx

managed by the ECM system as soon as it is created. An important part of Capture is the automatic classification of that content.

Manage: Manage is the component that provides the management of each individual content item through all its versions (lifecycle). It includes the following management applications:

- **Document Management (DM)**: A document in the context of ECM Document Management is defined as recorded information or an object which can be treated as a unit [1]. Today, this also includes E-Mail Management (EMM) and Digital Assets Management (DAM)[6].
- **Collaboration Tools (Collab)**: Collab includes the joint use and control over the content (including access management), as well as the applications that support this.
- **Web Content Management (WCM)**: Web Content Management is often controlled through what is now called a Content Management System (CMS) and may or may not be directly integrated with the ECM. Many ECM solutions provide access to the content via web-based user interfaces and many include actual WCM functionality.
- **Records Management (RM)**: Records Management is the management of what in this context is called Records. Records are content which will not change further and which, for legal reasons or because it may be of further relevance to the enterprise, must be stored for future reference [8]. In enterprise as well as in government environments, (electronic) records management (ERM) is governed by ISO standard 15489 [5].
- **Business Process Management (BPM) / Workflow (WF)**: BPM is a methodology to make processes efficient and effective by developing, deploying, monitoring, and optimising process automation applications. WF, as opposed to BPM, is the manual processes of managing documents in cases where human intervention is required (e.g. approval and prioritisation).

Store: The Store component of an ECM includes the actual physical locations (e.g. hard disks, storage area networks (SANs), or even CDs/DVDs) where the content is stored, as well as the logical structure of these physical locations. That structure, referred to as 'repository', can be a simple file system, a database or even a 'data warehouse'. Store also includes access strategies, also called 'library services', which include controlled check-in and check-out of content, search and retrieval mechanisms, version control, and the audit trail of each individual item. As such, it has a significant overlap with Manage.

Deliver (Output management): As search and retrieval (pull access) strategies are already controlled in the Manage and Store components of this model, Deliver is not concerned with this aspect. Deliver focuses on the control of external access to, and publication and distribution (push access) of content. This includes transformation of content for external access (e.g. text documents into personalised serial e-mails or letters or into PDFs for web publication), but also compression of files for storage or transformation of e.g. text documents to PDF/A [6] files for Preserve.

Preserve: The Preserve component deals chiefly with content that has been identified as Records by the Manage component. It is, obviously, directly related to the Store component as it deals with safe, long-term storage and back-up strategies (bit-stream preservation) for these Records. Preserve is fed either directly by Records Management and Manage or indirectly via Deliver, when content has been transformed for archiving.

2.2. ECM Compliance, Records Management, Preserve and Long-term Archiving

There are two main reasons why enterprise content that is no longer in regular use is preserved:

The first is known as Compliance. According to AIIM Compliance "means ensuring that the proper business practices are followed and that content is properly captured, stored, managed, and disposed of at the appropriate and legal time in its lifecycle."[3] This lifecycle may last for 10 years or more, during which time the respective content may need to be accessed, destroyed[7] or passed on to a different organisation (e.g. a national archive) at request or at a given time. It may be important that the business is able to prove the proper and legal destruction of said content, which will only be possible if that content is still accessible at the time it is to be destroyed. The second reason is more directly business related: Content may contain information about previous developments or projects, and this information may be of importance to later developments or projects.

Content that is no longer in active use but kept for either of the above mentioned reasons is referred to as Records. There is an ISO standard that regulates the procedures in handling Records. According to ISO 15489, a Record is "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" and (Electronic) Records Management ((E)RM) is the "field of management responsible for the efficient and systematic control of the creation, receipt, maintenance, use and

[6] DAM deals with any digital content that cannot be classified as a document.

[7] This is an important point in the comparison as the OAIS Model does not support the destruction of content.

disposition of records, including processes for capturing and maintaining evidence of and information about business activities and transactions in the form of records."[5]

As has been said above, Records are found in two of the main ECM components. The Manage component is responsible for deciding which Content will be kept as Records, while the Preserve component is responsible for actually archiving these Records. It is important to note that this mainly involves the storage of the Records. Migration in this context is still largely the migration of data from one storage medium to another and not, generally, the migration of content to different file-formats when the original format becomes obsolete. Content will only be migrated when passes through Deliver. This happens rather as a general strategy than as an actual act of digital preservation. If the format is chosen well (e.g. PDF/A for text documents) this may, however, have a similar effect for the accessibility of the content.

AIIM describes their approximation of Long-Term Archiving as "content that must be preserved over decades must be saved to media, such as paper and film-based imaging, with longevity to match."[2] There is some discussion about the 'transformation' of content into file formats that are preferable for long-term preservation. AIIM itself takes part in the development of the PDF/A standard. There are also considerations in the ECM and the ERM community about the transformation of e.g. CAD files into TIFF, JPEG or JPEG2000 files. [9] describes an example to keep CAD files usable, i.e. re-usable in a later project by migrating them an active CAD file-format, as it is impossible to produce e.g. a new architectural plan from static image files. These transformations are handled by Deliver and are thus connected to that part of the ECM system which usually deals with push-access and delivery of content, not with its preservation. No mention is made of management of long-term preservation that includes actual preservation planning as envisioned in the OAIS Model.[8]

3. THE OAIS MODEL

The OAIS Reference Model was first published in 1995, when the partners of the Consultative Committee for Space Data Systems realised that large portions of their data were no longer accessible due to changes in software and hardware systems. The model has been continuously refined. In January 2002 the OAIS Reference Model was published as a CCSDS Blue Book and has subsequently been adopted as an ISO Standard (ISO 14721:2003).

Reference Model standards, like OAIS, are developed in an open, public process. As the problems this standard addresses have become important beyond the space communities, the CCSDS set out to ensure broad participation from other fields – notably from the traditional archive community.[9] Since space data was no longer the only subject of the resulting model, the term Information was used to identify the content that would then be represented by the data in the archive.

3.1. Description

In the OAIS Reference Model, the Archival Information System includes hardware and software components as well as the people who are responsible for the acquisition, preservation and dissemination of the information. Additionally, the Model is designed as a framework for understanding, applying and discussing concepts needed for long-term digital preservation of information. Long-term, in this framework, means "long enough to be concerned about changing technologies."

3.1.1. OAIS Information and Information Package Definitions

One of the most important concepts in the OAIS Reference Model is that of **Information**. Information is defined as "any type of knowledge that can be exchanged." This Information is always represented as **Data**; and each individual instance of such Information is identified as an **Information Object**.

In order for an Information Object to be successfully preserved, it is necessary for the OAIS to clearly identify and understand the **Data Object** (the Data associated with that instance), and its associated **Representation Information**. The Representation Information is additional information that maps a Data Object into more meaningful concepts. Only in this combination is the Data Object usable and becomes the Information Object (or the object that was to be preserved). Without the Representation Information, the Data Object is often useless (see Figure 2).

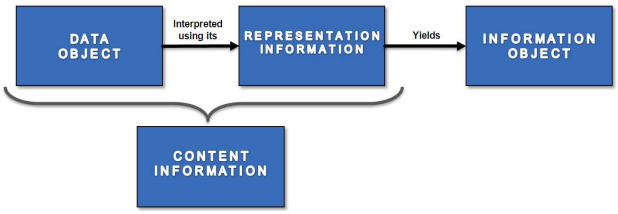


Figure 2. Relationship between Data Object and Information Object

Closely related to the concept of Information is that of the Information Package. An Information Package is a conceptual container of two types of Information called

[8] Kampffmeyer mentions the OAIS Model in [7] as a standard related to migration, but does not, apparently, consider the importance of preservation planning for ERM.

[9] The "Open" in OAIS is meant to signify this aspect of the modelling process and does not imply Open Access to the OAIS's content.

Content Information (the combination of Data Object and its associated Representation Information) and **Preservation Description Information**.

Preservation Description Information is the Information which is necessary for adequate preservation of the Content Information. It contains the following information:

Provenance: Provenance describes the history of the Content Information: Where it originated, what was changed (e.g. necessary format changes), who had custody of it since creation.

Reference: Reference identifies the Content Information (similar to an ISBN for a book).

Fixity: Fixity provides the authentication mechanisms and authentication keys to ensure that the Information Object has not been altered in an undocumented manner. This function is closely related to the concept of archival Authenticity, which is also relevant in Records Management and Compliance.

Context Information: Context Information documents the relationships of the Content Information to its environment.

3.1.2. OAIS Roles

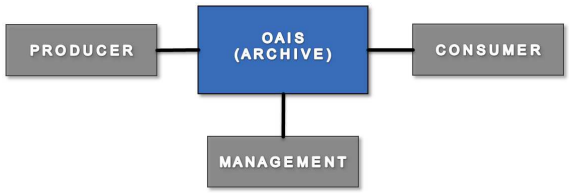


Figure 3. Simple OAIS Model

The simplest view of an OAIS (as shown in Figure 3) has three major roles attached to it:

Producer: Producer is the role of the entities (persons or client systems) that provide the Information to be preserved in the OAIS.

Management: Management is the role of those entities that set overall OAIS policy. These will usually have further management functions in the organisation the OAIS belongs to.

Consumer: Consumer is the role of those entities (persons or client systems) that interact with OAIS services to search for and access preserved Information.

An important OAIS concept related to the Consumer is that of the **Designated Community**. This is an identified group of potential Consumers of the OAIS. The Information to be preserved should be **Independently Understandable**. This means that it must be documented in such a way that any member of the Designated Community can understand it without external resources. The need to achieve this informs the decision on the content of the Representation Information. The broader the future community of potential Consumers is to be, the broader the content of the Representation Information must be.

3.1.3. OAIS Functional Entities and Data Flow

The following is an explanation of the functional entities of an OAIS (shown in Figure 4) and follows the flow of Data through the OAIS:

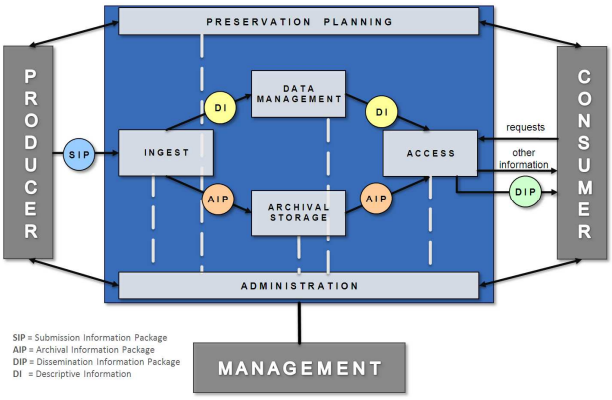


Figure 4: OAIS Functional Entities

Ingest: Ingest provides the services and functions for the OAIS to accept Information Packages from the Producers. These Packages are called **Submission Information Packages** (SIPs). The delivery of a SIP is negotiated between the Producer and the OAIS. At this point the OAIS assumes sufficient control of the Information to ensure long-term preservation. This means that it reserves for itself the right to manipulate the SIPs in such a way that they can be preserved. The OAIS also ensures that the Information is Independently Understandable to Designated Community by associating adequate Representation Information to it at this point. The OAIS follows documented policies and procedures (Preservation Plans), which ensure that the Information is preserved against all reasonable contingencies (e.g. through migration to another format). At Ingest, the content of the SIP is prepared for storage and management within the archive. Preservation Description Information is added to the Information Packages. The resulting **Archival Information Packages** (AIPs) are transferred to Archival Storage. The associated **Descriptive Information** (DI), which consists primarily of package descriptions, is provided to Data Management to support Access (the finding, ordering, and retrieving of OAIS Information holdings by Consumers).

Archival Storage: Archival Storage provides the services and functions for the storage, maintenance and retrieval of AIPs. It organises refreshing of storage media in order to provide the capability to reproduce the archive holdings over time (*bit-stream preservation*). For disaster recovery, Archival Storage provides a mechanism for producing duplicate copies of the AIPs in

the archive collection. Finally, Archival Storage provides copies of stored AIPs to Access.

Data Management: Data Management is the entity which provides services and functions for populating, maintaining, and accessing both DI and internal archive administrative data. It receives query requests from Access and generates result sets that are transmitted back to requesting Consumers. If the requested Data is available, Data Management generates a dissemination request which is sent to Access.

Access: Access supports Consumers in determining the existence, description, location and availability of Information stored in the OAIS and allows them to request **Dissemination Information Packages** (DIPs). A DIP is derived from part or all of one or more AIPs and is the Information Package that is sent to a Consumer. Among Access' functions are the finding aids, tools that provide an overview of the Information available in the OAIS.

Administration: Administration is the entity that manages the overall operation of the OAIS. Administration negotiates submission agreements with the Producers, manages system configuration, and develops the standards and policies for the OAIS. These include format standards, documentation standards, and the procedures to be followed during Ingest as well as the policies for storage management. Administration is responsible for Preservation Planning and for the audit of AIPs. The audit process must verify that the quality of the Data meets the requirements of the archive.

Preservation Planning: Preservation Planning is an important task of Administration. Preservation Planning interacts with Consumers and Producers to monitor changes in their respective service requirements and available technologies. Such requirements may include data formats, media choices, preferences for software packages or computing platforms, and available mechanisms for communicating with the OAIS (e.g. new finding aids for Consumers or ftp-up-load rather than SIP delivery by optical media for Producers). Preservation Planning is also responsible for tracking emerging digital technologies, information standards, and computing platforms (i.e. hardware and software), to identify technologies which could cause obsolescence in the OAIS's computing environment and thus loss of access to certain parts of the archive's holdings.

Internally, Preservation Planning develops packaging designs and detailed migration plans in order to implement Administration policies and directives. Preservation Planning receives approved standards and migration goals from Administration and implements these. Migration goals usually involve transformations of AIPs including, at times, transformations of the Content. Once the migration plan, associated AIP designs, and software have been tested and approved, the entire migration package is sent to Administration, which will execute the actual migration.

It is important to note that migration is not the only way to mitigate technology obsolescence. Other options include the emulation of obsolete hard- and software environments.

4. GAP ANALYSIS

From the descriptions in the previous sections can be seen that ECMs and OAISs have similar requirements in many areas. Other areas are only present either in one or in the other kind of system. One main difference, however, is organisational. While OAISs are conceived to be external organisations[10] that are independent of the creation process of the information they contain, ECMs actually facilitate such creation and control the whole lifecycle of the content. Thus, there is the Capture process that is designed to draw all creation of Information (or Content) of an enterprise into the ECM as a central point and as a first step to manage further versions and variations of that Information from within the ECM. Only when that creation process is over does the Content turn into Records and is handed over to ERM and Preserve. Most of the functionality that both ERM/Preserve and OAIS need (e.g. search and access) is provided within the ECM via Manage and Store. Certain provisions that are central to the OAIS Ingest function are also handled in an ECM system. An example of this is the creation of Metadata and Descriptive Information, which is already provided with the capture process and is maintained over the content's lifecycle.

Some ERM requirements are quite similar to those many archives or scientific organisations have. Archives live by rules that are not much different from those referred to by the term Compliance. In fact, the "Model Requirements Specification for the Management of Electronic Records" (MoReq2[11]) was produced at the request of the DLM Forum[12], which is an independent European community of both public archives and other organisations which deal in archiving, and records and information management. The need of a business organisation to preserve its internal information for future reference, on the other hand, may be likened to that of large scientific organisations, e.g. the CCSDS.

Organisations like AIIM have only recently recognised the fact that digital files have the inherent risk of becoming obsolete. This missing awareness may have been due to the relatively smaller size of the relative organisations' data archives, or to the differences in the length time that is usually envisioned for data

[10] This can also be independent parts of the organisations that produce the information.

[11] The MoReq2 (http://www.moreq2.eu/) specification has been prepared for the European Commission with funding from the IDABC (http://ec.europa.eu/idabc/) programme.

[12] http://www.dlmforum.eu/. The acronym "DLM" means "Document Lifecycle Management".

preservation. Only in recent years have people in the field become concerned with more than just the hardware side of preservation. AIIM is now a partner in the creation of the PDF/A standard, and several people in the field suggest image-file formats like TIFF, JPEG and more recently JPEG2000 for preservation. This can, however, only be a first step towards the integration of logical preservation into the ERM field.

As the number of formats used increases regularly, and more and more information is contained in a mix of formats (e.g. entire web-sites that contain sound and movie files besides the image and text files covered by the above mentioned suggestions), it is important that a culture of long-term digital preservation arises.

Such a culture has existed for a number of years in the scientific and the cultural memory fields, and the OAIS Model is its expression.

4.1. What is Required to Make an ECM OAIS Compliant?

As has been described above, an actual OAIS is made up of the following functional entities: Ingest, Archival Storage, Data Management, Access, Administration, and Preservation Planning. The Access functionality is already provided in the ECM through Capture, and different parts of Manage and Deliver, as is some of the functionality of Ingest, Archival Storage, and Data Management. What is totally missing is the combination of Administration and Preservation Planning. Together, these set the framework for proper digital preservation. They also provide control mechanisms and standards, and prescribe preservation policies, which set rules about when a certain type of preservation action is to be used on an endangered format. Preservation Planning provides the technology watch function (usually via external databases or technical registries[13]), which provides Administration with the triggers for such preservation actions. If, for example the external registry indicates that a particular format is at risk of no longer being supported by any software, Administration uses that indication to determine that now is the right time to migrate (in ECM terms transform) all files of that format. In this case, it is also the Preservation Planning component's responsibility to provide the plan on how (e.g. with which software, using which parameters etc.) and to which new format the files are to be migrated.

One important part of the OAIS Model is the description of the data flow within an OAIS. This not only ascribes responsibility to OAIS functionalities for the content at different stages of that data flow, but defines the additional information the OAIS must provide in addition to the original content. As has been shown above, ECMs do have similar functionality that provides e.g. version control and originator information (Provenance in OAIS terms) for content, but once the content turns into Records, no further - and more specifically no preservation-related – information is added. This means that large parts of the Preservation Description Information are missing.

In an OAIS, Representation Information and Preservation Description Information are added to the SIP at Ingest. Part of this is the identification of the proper file formats (i.e. with which version of a program was the file created, is it really the type of file the extension (e.g. '.doc') indicates, etc.). This again is usually handled by external services.[14] These steps are important, as only properly identified files can be successfully migrated according to the regulations of Preservation Planning. At present, no such information is added to the ECM Record.

Some of the information added to the SIP is provided to make its content Independently Understandable to the Designated Community. It can be assumed that for many ECMs the Designated Community's knowledge is equal to that of the Provider. It can perhaps be argued that in these cases no such information needs to be added, but a specific analysis of the circumstances may be advisable. Potentially, properly collected Preservation Description Information may be used as a source for Knowledge Management, thus ensuring that knowledge about certain processes and content are not lost to an enterprise when important employees leave or processes change.

As has been said earlier, Store, Manage and Deliver in ECMs provide much of the functionality that Archival Storage, Data Management and Access do in OAIS. It is, however, important that the manner in which these latter components work together is different. After Ingest, the AIP and its corresponding DI are sent to Archival Storage and Data Management respectively. Both are refreshed for each applied migration action. Following a request from Access, parts of the AIP and its DI are recombined to form the DIP. This functionality is, of course, not present in ECMs.

Additionally ECM Manage fulfils certain requirements that are part of an OAIS's Administration function. However, all requirements that deal with the audit of AIPs and all aspects of standard and policy development are missing.

So, while certain components of ECMs fulfil requirements of OAIS components, none of them completely do so, because the main purpose of an ECM is to provide an environment that allows for the active manipulation of content while the main purpose of an OAIS is to preserve Records (the ingested Information). The addition of such OAIS functionality to Records Management and Preserve can ensure that the information in the Records stays as close to the ingested

[13] E.g. the PRONOM technical registry. (http://www.nationalarchives.gov.uk/pronom/)

[14] E.g. the JSTOR/Harvard Object Validation Environment (JHOVE). (http://hul.harvard.edu/jhove/).

original as is possible and that additional information is kept to provide authenticity to the original content.

It appears therefore advisable to turn ERMs into proper OAISs while using functionality that already exists in other parts of the ECM.

One important function of ERMs that is not foreseen in the OAIS Model is that of the – legally required – destruction of certain Records after a given period of time. This function would of course have to be implemented into the new system.

5. SUMMARY

In this paper we have described and compared a general model for Enterprise Content Management systems and the Reference Model for an Open Archival Information System in order to determine what would be needed to make an ECM system OAIS-compliant and thus long-term preservation ready.

We have seen that some of the functionality needed for an OAIS can already be found in ECMs, but much of what it needs to become truly long-term reliable is missing:

1. Where ECM-Capture collects all content an organisation produces, OAIS-Ingest needs to be provided with the information that is to be preserved.
2. ECMs are usually integrated into the organisational infrastructure whereas OAISes are often external organisations that take responsibility for the preservation of the information other organisations have produced.
3. Capture collects metadata about ownership, access rights, and other information needed for the active part of a document's lifecycle. Ingest, while also responsible for some of these, specialises in preservation related metadata: e.g. file formats, representation and preservation metadata.
4. In an OAIS, descriptive information is held separately from the actual data (which represents the information that is to be preserved).
5. ECMs provide no Preservation Planning, watch functionality or controlled continuous logical preservation. To the OAIS these are provided by Administration.

One important point that can be seen from this analysis is that ECM and OAIS do not contradict each other; the OAIS functionality supplements that of the ECM. It offers missing functionality which may become crucial in providing businesses with Content Management Systems that assure Compliance for their digital assets.

6. REFERENCES

[1] Association for Information and Image Management (AIIM) *What is Document Management (DMS)?* AIIM., http://www.aiim.org/What-is-Document-Management-Systems-DMS.aspx. (Last accessed on 2010-03-23)

[2] Association for Information and Image Management (AIIM) *What is ECM?* AIIM. http://www.aiim.org/What-is-ECM-Enterprise-Content-Management.aspx. (Last accessed on 2010-03-23)

[3] Association for Information and Image Management (AIIM) *What is Electronic Records Management?* AIIM. http://www.aiim.org/What-is-ERM-Electronic-Records-Management.aspx. (Last accessed on 2010-03-23)

[4] Consultative Committee for Space Data Systems (CCSDS) *Reference Model for an Open Archival Information System (OAIS)*. CCSDS, 2002. http://public.ccsds.org/publications/archive/650x0b1.pdf. (Last accessed on 2010-03-23)

[5] ISO *ISO Standard 15489 (Information and documentation - Records management) Part 1: General*. ISO, Geneva, 2001.

[6] ISO *ISO Standart 19005-1(Document management - Electronic document file format for long-term preservation) - Part 1: Use of PDF 1.4 (PDF/A)*. ISO, Geneva, 2004.

[7] Kampffmeyer, U. et al *Effiziente Informationsverwaltung mit dem neuen europäischen Records-Management-Standard - MoReq2 und Records Management" (Seminarband)*. PROJECT CONSULT, Hamburg, 2008.

[8] Kampffmeyer, U. *ECM Enterprise Content Management – Whitepaper*. PROJECT CONSULT, Hamburg, 2006.

[9] Körber, N. *Dokumente rüsten für das Archiv* FEiG & PARTNER, Leipzig, 2006. http://www.documanager.de/magazin/artikel_1176_verwaltung_archivierung_digitaler_dokumente.html (Last accessed on 2010-03-06)

[10] Wikipedia *Enterprise Content Management*. Wikipedia, 2004. http://de.wikipedia.org/w/index.php?title=Datei:ECM_Komponenten.jpg (Last accessed on 2010-04-26)

FULL PAPER

A REFERENCE ARCHITECTURE FOR DIGITAL PRESERVATION

Gonçalo Antunes
INESC-ID
Rua Alves Redol 9, Apartado 13069, 1000-029 Lisboa, PORTUGAL

José Barateiro
LNEC
Av. Brasil 101, 1700-066 Lisboa, PORTUGAL

José Borbinha
INESC-ID
Rua Alves Redol 9, Apartado 13069, 1000-029 Lisboa, PORTUGAL

ABSTRACT

Apart from being a technological issue, digital preservation raises several organizational challenges. These challenges are starting to be addressed in the industrial design and e-Science domains, where emerging requirements cannot be addressed directly by OAIS. Thus, new approaches to design and assess digital preservation environments are required. We propose a Reference Architecture as a tool that can capture the essence of those emerging preservation environments and provide ways of developing and deploying preservation-enabled systems in organizations. This paper presents the main concepts from which a Reference Architecture for digital preservation can be built, along with an analysis of the environment surrounding a digital preservation system. We present a concrete Reference Architecture, consisting of a process to derive concrete digital preservation architectures, which is supported by an architecture framework for organizing architecture descriptions. In that way, organizations can be better prepared to cope with the present and future challenges of digital preservation.

1. INTRODUCTION

In order to achieve long-term digital preservation it is required to invest on a technical infrastructure for data storage, management, maintenance, etc. However, long-term digital preservation also raises several organizational challenges, since several business processes across the whole organization are affected by digital preservation.

Likewise, the complexity of long-term digital preservation increases with the fact that each type of business and specific organizations have their own particularities and special requirements, which makes the digital preservation business processes strongly dependent on their surrounding environment. For instance, the preservation policies depend on the type of data, its value for the organization, etc. As an example, the preservation of audio files requires recording information about compression and encoding/decoding which is not needed in the preservation of, for example, uncompressed XML files.

Concerning the organization type, memory institutions have several years of experience in dealing with the preservation of tangible objects. Additionally, the definition of preservation processes and policies concerning digital materials are common practices for these institutions. Usually, in the domain of memory institutions, technological solutions adopt the Reference Model for an Open Archival Information System (OAIS) [7], which provides a "framework for understanding significant relationships among the entities" involved in digital preservation. Actually, a framework can be described as "a set of assumptions, concepts, values, and practices that constitute a way of viewing the current environment" [12]. Reference frameworks can be used as basic conceptual structures to solve complex issues, providing a starting point to develop solutions concerning the targeted environment. Probably with the intention to support that, OAIS goes much further than providing just a high level reference model, detailing also on structural and behavioral issues.

Although the OAIS reference model has been widely adopted by memory institutions, it might not be suitable for scenarios with emergent digital preservation requirements, like industrial design. The OAIS reference model is definitely relevant for scenarios where the problem is to develop systems specifically for digital preservation, but it might not be appropriate for scenarios where the problem is to develop systems where digital preservation is a relevant property.

As a matter of fact, organizations with industrial design responsibilities produce a large amount of Computer-Aided Design (CAD) digital information within well-defined product lifecycles that cannot be aligned with the OAIS preservation processes and packages. Also, the collaborative environment of the scientific community, and associated services and infrastructures, usually known as e-Science (or enhanced

Science) [11], involves digital preservation requirements. Actually, long-term digital preservation can be thought as a required property for future science and engineering, to assure that information that is understood today is transmitted to an unknown system in the future.

In fact, we should recognize that, in the scope of digital preservation, it is crucial to better consolidate the perspective of the engineer (responsible for specific design and deployment of technological systems) to the perspective of the business architect (responsible by the business specifications, considering the related multiple systems, processes, and roles). Those concerns are already addressed by the Enterprise Architecture [1].

According to [1], a Reference Architecture "captures the essence of existing architectures and the vision of future needs and evolution to provide guidance to assist in developing new system architectures". In that sense, we intend to demonstrate that a Reference Architecture should not be an artifact, but a process from which multiple architectural artifacts can result and be governed throughout their lifecycle. Based on that, we propose a Reference Architecture for digital preservation, capturing the essence of preservation architectures so that system architectures that are preservation-enabled can be developed and deployed in organizations.

The motivation for this work comes from the national funded project GRITO[1] and the European funded project SHAMAN[2], where requirements for digital preservation in e-Science and Industrial Design are being addressed.

This paper is organized as follows. First, Section 2 describes the concepts of architecture, reference architecture, stakeholder, view, viewpoint and enterprise architecture. Second, Section 3 describes the digital preservation environment where a preservation system inhabits. Next, Section 0 presents a framework to support the Reference Architecture. Section 5 presents the Reference Architecture which consists of a process for the development of concrete preservation-enabled architectures. Finally, Section 6 presents the main conclusions and future work.

2. MAIN CONCEPTS

This section describes the main concepts of concerning Reference Architectures. These concepts have been derived from international standards and related models of the area.

2.1. About Architecture

According to the IEEE Std. 1471-2000[3], architecture is "the fundamental organization of a system, embodied in its components, their relationships to each other and the environment, and the principles governing its design and evolution" [8].

The standard describes that a system (which has a mission) inhabits an environment which influences it. The system has an architecture which is described by an architecture description, providing a rationale for the architecture. The architecture description identifies the stakeholders of the system, which have concerns about the system. For its turn, an architecture description may be composed of several views (which might include several models of the architecture), which are according to the viewpoint of the stakeholder (which is used to cover the concerns of the stakeholder). The viewpoints might originate from a viewpoint library. The concepts of Stakeholder, Viewpoint, and View will be described in the following sub-sections.

2.2. About Reference Architecture

A reference architecture [8] is a way of documenting good architectural design practices to address a commonly occurring problem. It is way of recording a specific body of knowledge, with the purpose of making it available for further practical reuse.

A relevant source to better explain and understand these concepts is the work of the Service Oriented Architecture (SOA) Technical Group from the Organization for the Advancement of Structured Information Standards (OASIS). According to their SOA Reference Model [12], "Concrete architectures arise from a combination of reference architectures, architectural patterns and additional requirements, including those imposed by technology environments."

Architecture must account for the goals, motivation, and requirements that define the actual problems being addressed. While reference architectures can form the basis of classes of solutions, concrete architectures will define specific solution approaches.

Architecture is often developed in the context of a pre-defined environment, such as the protocols, profiles, specifications, and standards that are pertinent. SOA implementations combine all of these elements, from the more generic architectural principles and infrastructure to the specifics that define the current needs, and represent specific implementations that will be built and used in an operational environment."

Therefore, reference architectures are relevant to support the development of specific concrete architectures.

[1] http://grito.intraneia.com/ (FCT, GRID/GRI/ 81872/2006)

[2] http://shaman-ip.eu/ (European Commission, ICT-216736)

[3] IEEE Std. 1471-2000 consists in a standard for the architectural description and design of systems, recommended by the IEEE Computer Society. http://www.computer.org/standards

2.3. About Stakeholders

A successful architecture has to reflect the concerns and interests of the stakeholders. In [13], architecture is described as "a vehicle for communication and negotiation among stakeholders". Taking that into account, the architecture must also reflect the different viewpoints of all the interested parts, so that it can be communicated efficiently.

Also in [13], a stakeholder is defined as a viewer that perceives and conceives the universe, using his/her senses, in order to produce conceptions resulting from the interpretation of what is observed. A viewer can form a representation of the conceptions he/she makes using a determined language to express himself. When observing the universe, a viewer will be interested only in a specific subset of that universe, which is called a concern. The conceptualization of that subset of the universe is called a domain.

The process of abstracting a domain in a model is called modeling. In order to start a modeling process, a viewer must first construct a meta-model, comprising the meta-concepts and modeling approach, when modeling a domain. Figure 1 depicts a generic situation where a viewer with a determined concern and meta-model conceives and represents models for several domains.

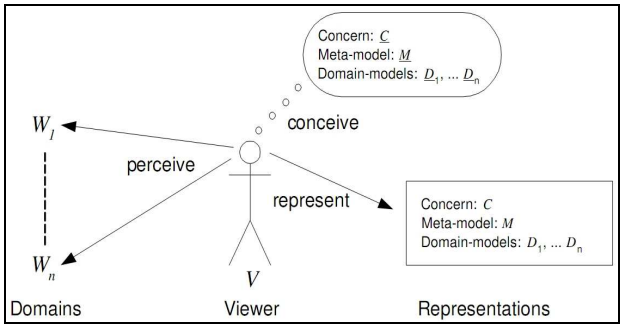


Figure 1. Viewing domains from a particular concern and meta-model [13].

Concluding, the concept of stakeholder has a crucial role in the development of an architecture since in order to be complete, an architecture should represent the different conceptions of the system through the use of models developed according to each of the relevant classes of stakeholders.

2.4. About Viewpoints and Views

Fundamental to the development of an architecture, and therefore to any reference architecture, are the concepts of "viewpoint" and of "view".

The concepts are distinct and the need for this distinction is justified since a viewpoint is a "formalization of groupings of models" through a template or pattern for representing a set of concerns of a stakeholder [8]. A view is the concrete representation of a entire system from the perspective of a viewpoint, through a set of models. The viewpoint provides the categorization and the view provides the models according to the categorization.

In order to be complete, an architecture description must be composed of multiple views, addressing the concerns of multiple stakeholders. About the use of multiple views, the standard considers the following [8]: "The use of multiple views to describe an architecture is therefore a fundamental element of this recommended practice. However, while the use of multiple views is widespread, authors differ on what views are needed and on appropriate methods for expressing each view". Although the standard does not prescribe a set of views or modeling techniques for developing views, the field of Enterprise Architecture provides some examples of the views that should be considered in an architecture description.

2.5. About Enterprise Architecture

Enterprise Architecture is defined as a coherent whole of principles, methods, and models that are used in the design and realization of an enterprise's organizational structure, business processes, information systems, and infrastructure [10]. An Enterprise Architecture framework is a communication tool to support the Enterprise Architecture process. It consists of a set of concepts that must be used as a guide during that process.

One of the first Enterprise Architecture frameworks was the Zachman framework [15], defined as "...a formal, highly structured, way of defining an enterprise's systems architecture. (...) to give an holistic view of the enterprise which is being modeled."

The Zachman framework is summarized in simple terms in Table 1, where each cell on the table can be related to a set of models, principles, services, standards, etc., whatever is needed to register and communicate its purpose.

The columns of the Zachman framework express the viewpoints relevant for this scope: the "What" refers to the system's content, or data; the "How" refers to the usage and functioning of the system, including processes and flows of control; the "Where" refers to the spatial elements and their relationships; the "Who" refers to the actors interacting with the system; the "When" represents the timing of the processes; and the "Why" represents the overall motivation, with the option to express rules for constraints where important for the final purpose.

The meaning of the rows are: "Scope" defines the business purpose and strategy; "Business Model" describes the organization, revealing which parts can be automated; "System Model" describes the outline of how the system will satisfy the organization's information needs, independently of any specific technology or production constraints; "Technology

Perspective / Role	DATA What	FUNCTION How	NETWORK Where	PEOPLE Who	TIME When	MOTIVATION Why
Planner (Objective/Scope - Contextual)	Things important for the business	Business Processes	Business Locations	Important Organizations	Events	Business Goals and Strategies
Owner (Enterprise Model – Conceptual)	Conceptual Data / Object Model	Business Process Model	Business Logistics System	Workflow Model	Master Schedule	Business Plan
Designer (System Model –Logical)	Logical Data Model	System Architecture Model	Distributed Systems Architecture	Human Interface Architecture	Processing Structure	Business Rule Model
Builder (Technology Model – Physical)	Physical Data/Class Model	Technology Design Architecture	Technology Architecture	Presentation Architecture	Control Structure	Rule Design
Programmer (Detailed Representation – Out of Context)	Data Definition	Program	Network Architecture	Security Architecture	Timing Definition	Rule Speculation
User (Functioning Enterprise)	Usable Data	Working Definition	Usable Network	Functioning Organization	Implemented Schedule	Working Strategy

Table 1. The Zachman Framework

Model" tells how the system will be implemented, with the specific technology and ways to address production constraints; "Components" details each of the system elements that need clarification before production; and "Instances" give a view of the functioning system in its operational environment.

The Zachman framework influenced many other Enterprise Architecture frameworks [3].One of those frameworks is The Open Group Architecture Framework (TOGAF), which consists of a "detailed method and a set of supporting tools" [14]. It is divided in seven parts, the most relevant being the Architecture Development Method (ADM), the Architecture Content Framework, and the Enterprise Continuum and Tools.

The ADM is defined as the core of TOGAF. It consists of a cyclical process divided in nine phases, which begins with the elaboration of the architecture principles and vision and goes through the elaboration of the concrete architectures and consequent implementation.

The Architecture Content Framework is TOGAF alternative to the use of the Zachman framework or any other architecture framework. The Content framework divides the types of architecture products in deliverables, artifacts and building blocks. Deliverables represent the output of the projects and are contractually specified. Artifacts describe architecture from a specific viewpoint, an example being a diagram. Building blocks are reusable components of business, IT, or architectural capability which can be combined to deliver architectures and solutions. Deliverables are composed of artifacts which for its turn describe building blocks. The Enterprise Continuum classifies the assets that may influence the development of concrete architectures. It contains two specializations, the Architecture Continuum and the Solutions Continuum. The Architecture Continuum classifies the architectures in Foundation Architectures, Common Systems Architectures, Industry Architectures, and Organization-Specific Architectures. These can be used to guide and support the development of Solutions, which the Solution Continuum classifies as Foundation Solutions, Common Systems Solutions, Industry Solutions, and Organization-Specific Solutions.

The Reference Architecture presented in this paper is largely inspired by TOGAF. It comprises an architectural framework and a process for the development of preservation architectures.

3. DIGITAL PRESERVATION ENVIRONMENT

As referred in Section 2.1, a "System inhabits an environment" which, for its turn, "influences the system".

Research undertaken in the SHAMAN project reached the conclusion that a bigger understanding of the environment where the preservation system operates is required [4]. A way of understanding the implications of the context of a digital object is through the analysis of its lifecycle. OAIS restricts itself to the "inner walls" of the archive, which may be insufficient in terms of the additional information required to preserve the object. A broader notion of the object lifecycle is needed, so that all the knowledge necessary to reuse the objects in the future is also preserved. The lifecycle of the digital object is represented in Figure 2.

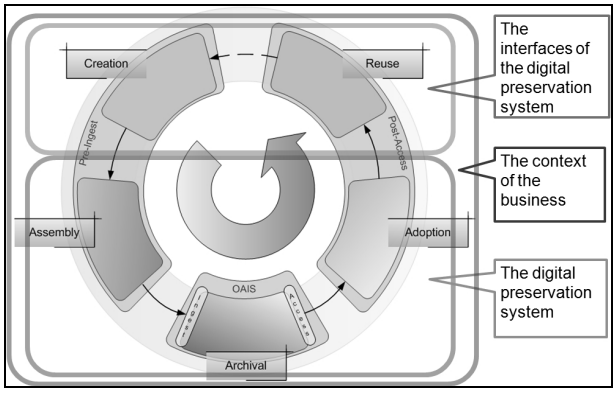


Figure 2. The Context of Digital Preservation in SHAMAN (adapted from [4]).

The **Archival** phase spans the OAIS scope. **Creation** is the initial phase during which new information comes into existence. **Assembly** denotes appraisal of objects relevant for archival and all processing and enrichment for compiling the complete information set to be sent into the future, meeting the presumed needs of the designated community. It requires deep knowledge about the designated community in order to determine objects relevant for long-term preservation together with the information about the objects required for identification and their reuse some time later in the future. **Adoption** encompasses all processes by which information provided by the Archive is screened, examined, adapted, and integrated for Reuse. This phase might comprise transformations, aggregations, contextualization, and other processing required for repurposing of data. **Reuse** means the exploitation of information in the interests of the consumer and other processing required for repurposing of data.

Taking all this into account, in the perspective of the SHAMAN project, the digital preservation system encompasses the phases comprised in the OAIS specification in addition to the Assembly and the Adoption of digital objects.

Considering the lifecycle of digital objects, the environment of the preservation system can be determined to be all that is outside and interfaces with the preservation system. In other words, the environment of the preservation system corresponds to the preservation "business" which the preservation system is supposed to support.

Taking into consideration this context of the preservation business and using Risk Management terminology [9], a taxonomy of threats and vulnerabilities of digital preservation, which takes technological, organizational, and contextual issues, can be devised [2].

Table 2 presents the taxonomy along with a classification of the threats and vulnerabilities according to the issues that may cause them (the capital characters represent bigger impact of a determined issue). The Reference Architecture for digital preservation draws from this analysis and is presented in the next sections.

Vulnerabilities	**Process**	Software faults Software obsolescence	T T	. .	. .
	Data	Media faults Media obsolescence	T T	. .	. .
	Infrastructure	Hardware faults Hardware obsolescence Communication faults Network service failures	T T T T	. . . .	. . . .
Threats	**Disasters**	Natural Disasters Human operational errors	T t	. O	C .
	Attacks	External attacks Internal attacks	t t	o O	C c
	Management	Organizational failures Economic failures	. .	O O	. c
	Business Requirements	Legal requirements Stakeholders' requirements	. .	. o	C C

Table 2. Taxonomy of Threats and Vulnerabilities to Digital Preservation

Stakeholders	Decision Making	Designated Community	Requirements and Conformance	Preservation Strategic Planning	**Viewpoints**
		Regulator			
		Auditor			
		Preservation Manager	Business Governance		
		Organization Manager			
		Technology Manager			
	System Building and Operation	System Designer	System Building and Support		
		Technology Provider			
		Technology Operator			
		Preservation Operator	Acting and Operation		
		Producer			
		Consumer			

Table 3. The Reference Architecture Framework

4. REFERENCE ARCHITECTURE FRAMEWORK

An architecture description identifies the stakeholders of the system and is composed of several viewpoints that reflect the concerns of the stakeholders [8]. In this section, we present a framework for architecture descriptions to support the Reference Architecture.

Following the guidelines of the IEEE Std. 1471-2000, the stakeholder identification should take into account [8]: (*i*) the users of the system; (*ii*) those responsible for the acquisition and governance of the system; (*iii*) the developers and providers of the system's technology; and (*iv*) the maintainers of the system as a technical operational entity.

4.1. Stakeholders

The classes of stakeholders identified upon to this moment are: (i) Designated Community - As stated in OAIS, this is "an identified group of potential consumers who should be able to understand a particular set of information. The Designated Community may be composed multiple user communities". It may affect the design and development of the preservation system, since the system should satisfy their requirements; (ii) Preservation Manager - The person responsible for the definition and management of preservation policies (but that does not operate with the system, as that is the role of the Preservation Operator); (iii) Regulator - The person responsible for any external imposing rules concerning the preservation business, such as legislation, standards, etc. Those can apply to the organization, the technology, or the systems' usage; (iv) Auditor - The person responsible for the auditing and certification of the organization compliance with the established standards, rules and regulations; (v) Organization Manager - The top of the organizational structure with the main responsibility of defining the overall business objectives and strategy. It is typically a Chief Executive Officer, but it also might be a committee; (vi) Technology Manager - The person responsible for the definition of the overall technological strategy (software, hardware and infrastructure in general). It is typically called a Chief Information Officer, but it also might be a committee; (vii) Consumer - Represents the user accessing to the preserved objects, with a potential interest in its reuse; (viii) Producer - The person responsible for the ingestion of the objects to be preserved (the owner of the object, but it also can be any other entity entitled for that); (ix) Preservation Operator - The business worker responsible for the operation of the system. It may be aware of the details of the design and deployment of the system, but its main concern must be to assure the direct support to the business; (x) System Designer - The person responsible for the design and update of the architecture of the system, aligned with the business objectives; (xi) Technology Provider - The person responsible for the implementation and deployment of the architecture of the system or only its components; and (xii) Technology Operator - The person responsible for the regular operation and maintenance of the technological infrastructure (user accounts, replacement of damaged components, etc.).

4.2. Viewpoints

After the analysis of the stakeholders and their concerns, the viewpoints listed in Table 3 were derived. The main source used for that was the *Trustworthy Repositories Audit and Certification: Criteria and Checklist* (TRAC) [5], due to its wide scope view.

These viewpoints are: (*i*) **Preservation Strategic Planning** - Deals with the organization process of defining the digital preservation mission, vision and strategy in the context of the organization-wide mission, vision and strategy. It defines the direction of the organization concerning preservation. Although generally elaborated by the top-level management, it concerns all the stakeholders; (*ii*) **Requirements and Conformance** - Deals with the extra-organizational context that influences the adoption or operation of the system. It might be at the level of requirements of potential users or at the level of the legal framework that regulates preservation activities, also including the auditing of the system and involved processes; (*iii*) **Business Governance** - Deals with the high-level management of the preservation infrastructure, in terms of regulation, policies, best-practices, etc. It comprises three level: organizational, preservation and technological; (*iv*) **Acting and Operation** - Deals with the usage of the system and all the administrating and operational tasks related to preservation; and (*v*) **System Building and Support** - Deals with the technical analysis, design, implementation, and deliver of the system or of its components, including the related infrastructure.

The viewpoints can be further divided in sub-viewpoints which will correspond to models of the architecture. Each of these sub-viewpoints will correspond to a model which can be developed using the Unified Modeling Language (UML), or other formal or informal representation technique. For example, a sub-viewpoint of the Preservation Strategic Planning viewpoint is the Preservation Principles Catalog, which contains a list of all the Preservation Principles that the architecture must comply with. The representation of this sub-viewpoint can be made through a table or a list.

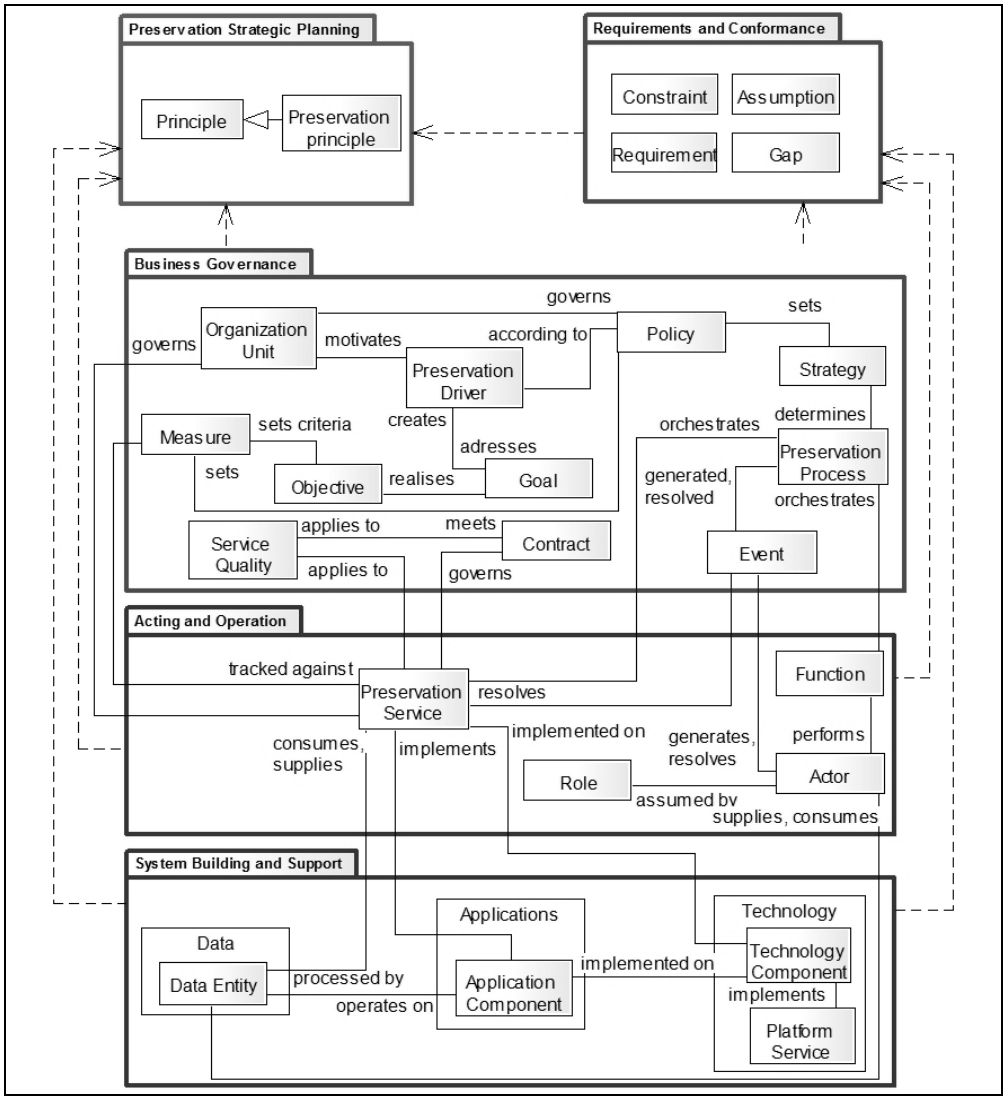


Figure 3. Reference Architecture Meta-model

4.3. Architecture Meta-model

The Architecture Meta-model provides a set of entities of the digital preservation domain, including the relationships between them. Those entities provide a common language for the domain which should be used on the development of the viewpoints of the architecture, when instantiating concrete architectures derived from the reference architecture. The meta-model enables the tracing between the different entities of the domain on the models of the architecture that result from the application of the Reference Architecture, enhancing the alignment between different viewpoints.

The meta-model is based in the TOGAF Content Meta-model of the Content Framework [14]. Figure 3 represents the entities of the digital preservation domain and relationships between the entities of the meta-model and also the relations between the viewpoints of the meta-model, using the Unified Modeling Language (UML).

5. REFERENCE ARCHITECTURE FOR DIGITAL PRESERVATION

A Reference Architecture "provides guidance to assist in developing new systems architecture" 8. In that sense, should be a process which origins and governs the lifecycle of architecture artifacts, supported by a framework, which was presented in the previous section.

The IEEE Std. 1471-2000 does not provide or recommends a methodology for architecture development [8]. In other hand, the TOGAF specification [14], which is aligned with the IEEE Std. 1471-2000, provides a solid and detailed method for the development of architectures. Therefore, it was decided to base the SHAMAN architecture development process in the principles of the TOGAF Architecture Development Method (ADM). The result was the SHAMAN Architecture Development Method (SHAMAN-ADM).

The SHAMAN-ADM comprises six different phases (Figure 4), which are in line with the architecture viewpoints of the reference architecture framework presented in Section 4.

The Preservation Strategic Planning phase deals with the initiation of the architectural activities, comprising the definition of the enterprise scope of the architecture, the existing organizational context, (preservation) business requirements, the architecture principles, the identification of the relationships between the architectural framework and other governance frameworks, evaluating the maturity of the architecture, and developing an Architecture Vision that provides guidance throughout the development of the architecture.

The Business Governance phase is concerned with the development of a business governance architecture for digital preservation that supports the Architecture Vision. The Acting and Operation phase determines the requirements and functions required by the actors of the system, supporting the Architecture Vision.

The System Building and Support is divided in three sub-phases. The Data Architecture phase determines the data needed to support the effective preservation of digital objects. Also, data migration requirements should be supported by the data architecture resulting from this phase. The Applications Architecture phase defines the applications needed to support the data and business of digital preservation. The Technology Architecture determines the technology components needed to support the application components defined in the previous phase. Finally, the Architecture Realization phase is concerned with the architecture implementation process.

The Requirements and Conformance should be a continuous practice throughout the application of the ADM. The management of requirements should be dynamic and preservation requirements at all levels shall be identified and stored, fed into and out of all the phases of the development cycle.

The application of this process in conjunction with the Reference Architecture framework should result in a architecture with preservation properties and in conformance with the requirements of the preservation stakeholders.

6. CONCLUSIONS AND FUTURE WORK

This paper presented a Reference Architecture for Digital Preservation. This work demonstrates a framework and a process from which concrete systems architectures with preservation properties can be derived, addressing particularly two digital preservation domains which introduced new and emergent requirements that cannot be addressed directly by OAIS: the Industrial Design and the e-Science domains.

We also presented the main concepts which form the background to the Reference Architecture, namely the concepts of Architecture, Reference Architecture, Stakeholder, Viewpoint and View, and Enterprise Architecture. Additionally, we motivated our approach through a general analysis of the digital preservation environment.

Future work will now focus on the application of the Reference Architecture to concrete cases to be explored on the scope of the SHAMAN project, which will result in the production of preservation-enabled architecture for specific cases. Another possible result may be a specialization of the reference architecture into the three domains of focus explored by the project, if irreconcilable differences are found between the domains.

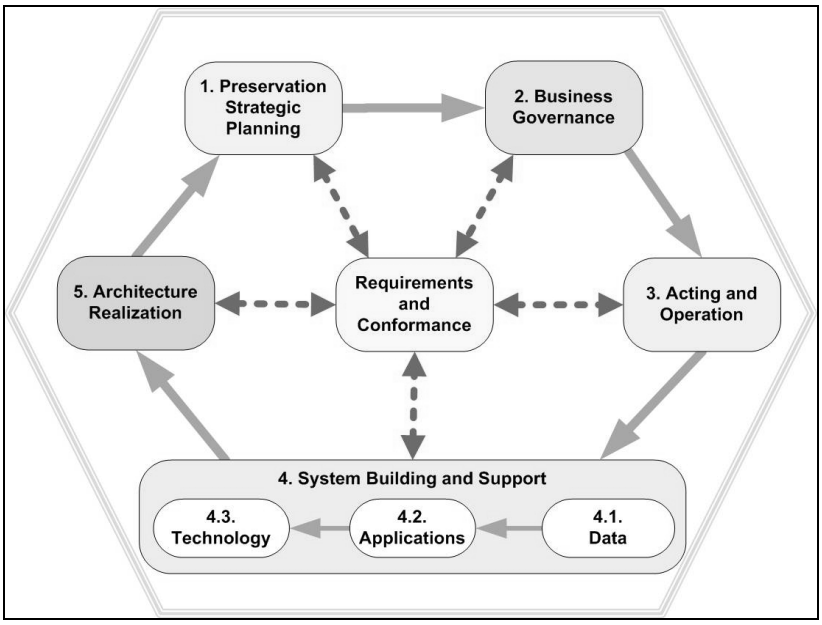


Figure 4. The Reference Architecture Development Method (SHAMAN-ADM).

7. ACKNOWLEDGEMENTS

The research reported was mainly supported by the project GRITO (a Grid for Digital Preservation), funded by the FCT (Portuguese Foundation for Science and Technology) under the contract GRID/GRI/81872/ 2006, and by the project SHAMAN (Sustaining Heritage Access through Multivalent Archiving), funded under 7th Framework Programme of the EU under the contract 216736.

8. REFERENCES

[1] Anaya, V., Ortiz, A. "How enterprise architectures can support integration", *Proceedings of the First International Workshop on Interoperability of Heterogeneous Information Systems*, Germany, 2005.

[2] Barateiro, J., Antunes, G., Freitas, F., Borbinha, J. "Designing digital preservation solutions: a Risk Management based approach", *In the 5th International Digital Curation Conference - "Moving to Multi-Scale Science: Managing Complexity and Diversity"*, London, UK, 2009.

[3] Borbinha, J. "It is Time for the Digital Library to Meet the Enterprise Architecture", *In Proceedings from the 10th International Conference on Asian Digital Libraries*, Hanoi, Vietnam, 2007.

[4] Brocks, H., Kranstedt A., Jäschke, G., Hemmje, M. "Modeling Context for Digital Preservation", *Smart Information and Knowledge Management: Advances, Challenges, and Critical Issues (Springer Studies in Computational Intelligence)*, 2010.

[5] Center for Research Libraries. *Trustworthy Repositories Audit and Certification (TRAC) Criteria and Checklist.* Chicago, 2008.

[6] Cloutier, R., Muller, G., Verma, D., Nilchiani, R., Hole, E., Bone, M. "The Concept of Reference Architectures", *Systems Engineering, 13* (1), pp. 14-27, Wiley Periodicals, 2010.

[7] Consultative Committee on Space Data Systems. *Reference model for an open archival information system.* ISO 14721:2003, 2003.

[8] IEEE Computer Society. *IEEE Std. 1471-2000: IEEE Recommended Practice for Architecture Description of Software-Intensive Systems*. IEEE, New York, 2000.

[9] International Organization for Standardization. *ISO 31000: Risk Management Principals and Guidelines*. ISO, Geneva, Switzerland, 2009.

[10] Lankhorst, M. *Enterprise Architecture at Work: Modelling, Communication, and Analysis*. Springer, Berlin/Heidelberg, 2005.

[11] Miles, S., Wong, S. C., Fang, W., Groth, P., Zauner, K.-P., Moreau, L. "Provenance based validation of e-science experiments", *Web Semant,* 5(1):28-38, 2007.

[12] Organization for the Advancement of Structured Information Standards. *Reference Model for Service Oriented Architecture version 1.0*. OASIS Standard, 2006.

[13] Proper, H. A., Verrijn-Stuart, A. A., Hoppenbrouwers, S. J. B. A. "On Utility-based Selection of Architecture-Modelling Concepts", *Proceedings of the 2nd Asia-Pacific conference on Conceptual Modelling*, Newcastle, NSW, Australia, 2005.

[14] The Open Group. *TOGAF Version 9*. Van Haren Publishing, Zaltbommel, Netherlands, 2009.

[15] Zachman, J. "A Framework for Information Systems Architecture", *IBM Systems Journal*, 12 (6), pp. 276-292, 1987.

FULL PAPER

Policy-Driven Repository Interoperability: Enabling Integration Patterns for iRODS and Fedora

David Pcolar
Carolina Digital Repository (CDR)
UNC Chapel Hill
david_pcolar@unc.edu

Daniel W. Davis
Cornell Information Sciences (CIS)
DuraSpace Affiliate
dwdavis@cs.cornell.edu

Bing Zhu
Data Intensive Cyber Environments (DICE)
University of California: San Diego
bizhu@ucsd.edu

Alexandra Chassanoff
School of Information & Library Science (SILS)
UNC Chapel Hill
achass@email.unc.edu

Chien-Yi Hou
Sustainable Archives & Leveraging Technologies (SALT)
UNC Chapel Hill
chienyi@unc.edu

Richard Marciano
Sustainable Archives & Leveraging Technologies (SALT)
UNC Chapel Hill
richard_marciano@unc.edu

ABSTRACT

Given the growing need for cross-repository integration to enable a trusted, scalable, open and distributed content infrastructure, this paper introduces the Policy-Driven Repository Interoperability (PoDRI) project investigating interoperability mechanisms between repositories at the policy level. Simply moving digital content from one repository to another may not capture the essential management policies needed to ensure its integrity and authenticity. Platform-independent, policy-aware object models, including policy expressions, and a distributed architecture for policy-driven management are fundamental building blocks of a sustainable access and preservation infrastructure. This project integrates iRODS and Fedora to demonstrate such an infrastructure. Using iRODS and its rules engine, combined with Fedora's rich semantic object model for digital objects, provides the basis for implementing a policy-driven test-bed. Using a policy-driven architecture is an essential part of realizing a fully model-driven repository infrastructure capable of decoupling the permanent digital content from the constantly evolving information technology used to support them.

1. INTRODUCTION

This paper introduces the Policy-Driven Repository Interoperability (PoDRI) project investigating interoperability between repositories at the policy level. PoDRI is led by the University of North Carolina at UNC, with units ranging from SALT (Sustainable Archives & Leveraging Technologies), RENCI (Renaissance Computing Institute), SILS (School of Information and Library Science), and the Libraries/CDR (Carolina Digital Repository). Key partners include Bing Zhu at UCSD (DICE, Data Intensive Cyber Environments) and Daniel Davis at DuraSpace (combining DSpace and Fedora Commons) and Cornell Information Sciences. The project is sponsored by an IMLS National Leadership grant and is motivated by the growing need to create a scalable, open and distributed infrastructure that provides durable, trusted access and management of our valuable digital content of all kinds (e.g. research data sets, documents, video, metadata). This problem is well described in the NSF's Cyberinfrastructure Vision for the 21st Century [14].

Simply replicating digital content from one repository, with or without any associated metadata, may not capture the essential management policies that ensure integrity and authenticity, a critical requirement for establishing a trust model. "A policy is typically a rule describing the interactions of actions that take place within the archive, or a constraint determining when and by whom an action may be taken [8]." Typical policies include those that control data ingestion, administration, preservation, access procedure, authentication and authorization.

A distributed policy management architecture is an essential component in realizing a trust mechanism for repository interoperability. The PoDRI project investigates the requirements for policy-aware interoperability and demonstrates key features needed for its implementation. The project is focused on integrating object models, including interoperable policy expressions, and a policy-aware distributed architecture that includes both repositories and middleware services.

Our overarching design paradigm is that permanent digital content must be decoupled from the constantly evolving infrastructure supporting it. Increasingly, the information infrastructure will be a part of a global, interoperable, heterogeneous, distributed "system-of-systems". Model-driven methods will be essential to make the management of such an infrastructure feasible; policy-driven methods are a core, enabling part of governance mechanisms needed to ensure control and preservation of our permanent digital content.

The PoDRI project addresses the following research problem: **What is the feasibility of repository interoperability at the policy level?** Research questions to be addressed are:

- Can a preservation environment be assembled from two or more existing repositories?
- Can the policies of the federation be enforced across repositories?
- Can policies be migrated between repositories?
- What fundamental mechanisms are needed within a repository to implement new policies?

iRODS (integrated Rule-Oriented Data System) [12,14] and the Fedora Repository [7,9] will be used as representative open source software to demonstrate the PoDRI architecture. Combining iRODS and Fedora enables use of the best features of both products for building sustainable digital repositories. iRODS provides an integrated rule engine, distributed virtual storage, the iCAT[1], and micro-services[2]. Fedora offers rich semantic object modeling for digital objects, extensible format-neutral metadata and a flexible service mediation mechanism.

2. RATIONAL FOR IRODS-FEDORA INTEGRATION

Early in 2006, the DART project [3] created an SRB storage interface for Fedora that allows all Fedora digital content, including Fedora Digital Objects (FDO) and their Datastreams, to be stored in SRB distributed repositories. Similarly, a storage module was developed by Aschenbrenner and Zhu [1] for iRODS. Using the Fedora-iRODS storage module, iRODS can act as a back-end for Fedora and, thus, provide opportunities for Fedora to use iRODS capabilities such as virtual federated storage, micro-services and the rules engine.

iRODS offers an appealing platform for implementing a distributed policy-driven management architecture. The integrated rules engine can be used to invoke a range of rules including policy expressions and, through the use of micro-services, can execute code for those policies in a distributed environment. Rules can act as simple workflows performing a sequence of pre-defined actions. iRODS rules can be executed explicitly, triggered by external conditions or events and executed at timed intervals. For example, iRODS can implement a replication policy, geographically disbursing file copies across the network. Micro-services can be written for feature extraction, format migration, integrity checks and other preservation services.

While used to efficiently hold and query structured data and metadata, the iCAT relational database is not optimal for handling the complex, variable metadata needed for preservation and curation. Indeed, any relational database will require considerable coding to support complex metadata schemas, making the use of unstructured data (files) possibly in combination with XML databases or semantic triplestores as a more flexible alternative [10].

Fedora is file-centric; all Fedora data and metadata is stored in files [6]. The Fedora Digital Object (FDO), a kind of compound digital object, provides the organizing metadata used to "make sense" of itself and other resources. It uses the FOXML schema to encapsulate metadata, and to reference other files or web resources. Since the FDO is a file, it can be stored in iRODS like any other file.

Digital content (or user-defined metadata) managed by the FDO is stored in one or more separate files — each registered in a FOXML element called a Datastream. Datastreams can also capture relationships to other objects and external resources. Users may add metadata to the FDO or add additional metadata Datastreams (to be stored like any other file).

This means, however, that metadata is stored in an unstructured format, often XML or RDF, and requires external indices to support querying by search engines, semantic triplestores, XML databases, and the iCAT. Fedora's approach provides a format neutral, extensible framework for representing data and metadata.

The rich metadata environment provided by the FDO can augment the structured metadata found in the iCAT. Metadata can be copied from the iCAT into a more easily preserved unstructured file format, as demonstrated by Bing Zhu and colleagues [17]. Critical data can be copied from the FDO, or as user metadata files (Datastreams), so they can be queried from the iCAT. With suitable metadata both the iCAT and the Fedora repository could be entirely rebuilt from files if the indices were lost or corrupted.

Fedora has a set of "front-end" APIs that provide the means to ingest and manipulate FDOs (CRUD). iRODS is capable of calling these APIs to perform operations from micro-services. Fedora also provides an extensible mechanism to add custom functionality called "services" that are executed within the context of the FDO. Services act as extensions to the "front-end" API of the object. Fedora mediates the service request calling the appropriate "back-end" functionality. The back-end functionality can be a Web service, in this case potentially provided by iRODS. Custom Fedora services provide another mechanism to interact with iRODS. Since iRODS can interact with Fedora's "front-end"

[1] iCAT is the metadata catalog in iRODS that stores metadata about all objects in iRODS in relational databases.

[2] Microservices are function snippets or executables that can be used to perform a distinct task using well-defined input information structures.

APIs, "back-end" services, and the Fedora-iRODS storage module, one may picture iRODS wrapping around Fedora.

3. ENABLING A POLICY-DRIVEN MANAGEMENT ARCHITECTURE

To demonstrate distributed policy-driven management architecture, we plan to implement the following operational scenarios:

- Integrate views of content, original arrangement (hierarchy) and metadata
- Create an audit trail of policy execution events and related provenance information
- Manage policies through Fedora
- Show iRODS invoking policies from Fedora

Both iRODS and Fedora fully support distributed computing installations. In effect, both products can be characterized as virtualization middleware for storage, access and service execution. The products, however, have very different operational paradigms which must be accommodated but provide complementary strengths that can be exploited when used together.

The virtual file system in iRODS makes it the logical choice for all storage (including FDOs). In addition, the iRODS rules engine and micro-services provide an effective means for orchestrating services such as policy invocation. Fedora's capabilities, on the other hand, are especially powerful for handling variable content and different metadata formats, for flexibly relating resources, facilitating presentation (manifestation of content), and its mediation capabilities make it appealing in building systems that are "designed for change."

A policy-driven management architecture requires that policy expressions be persistent. Fedora could be used to create FDOs containing policy expressions, which would subsequently be loaded into machine-actionable form and invoked as iRODS rules. Since policies are part of an object's provenance, Fedora can relate the policy FDOs to content items in which they apply. Because policy invocation will be performed by iRODS, audit records of the execution must also be created by iRODS. Subsequently, iRODS will store the execution records back into Fedora as FDOs, linking them to the FDOs containing the content and policy expressions.

iRODS does not currently generate audit data in a format compliant with the PREMIS preservation metadata schema. The CDR, however, implements auditing of objects via a PREMIS.XML file for each iRODS data object. This method may not be sustainable for repositories containing millions of objects. Preservation activities, such as replication or fixity checks, generate large amounts of log entries over time and potentially exceed the byte size of the original object. Discussions between CDR and iRODS developers suggest multiple methods for retaining and aggregating various component logs for translation into PREMIS-compliant events. Do we continue to store these events with the individual objects or as an aggregate? Do we generate specific PREMIS information upon request? In the case of replicas residing on disparate nodes in a data grid, auditable events will occur that differ from those affecting the original object. How do we reconcile these events in a singular view of the object?

Users and user applications will still need to interact with Fedora or iRODS directly. This is particularly true of research (grid) applications with large datasets. Select metadata will need to be duplicated in both products to access content, to represent relationships, and to preserve integrity and authenticity. Direct interaction by users or user applications with either Fedora or iRODS will require both products to synchronize or update

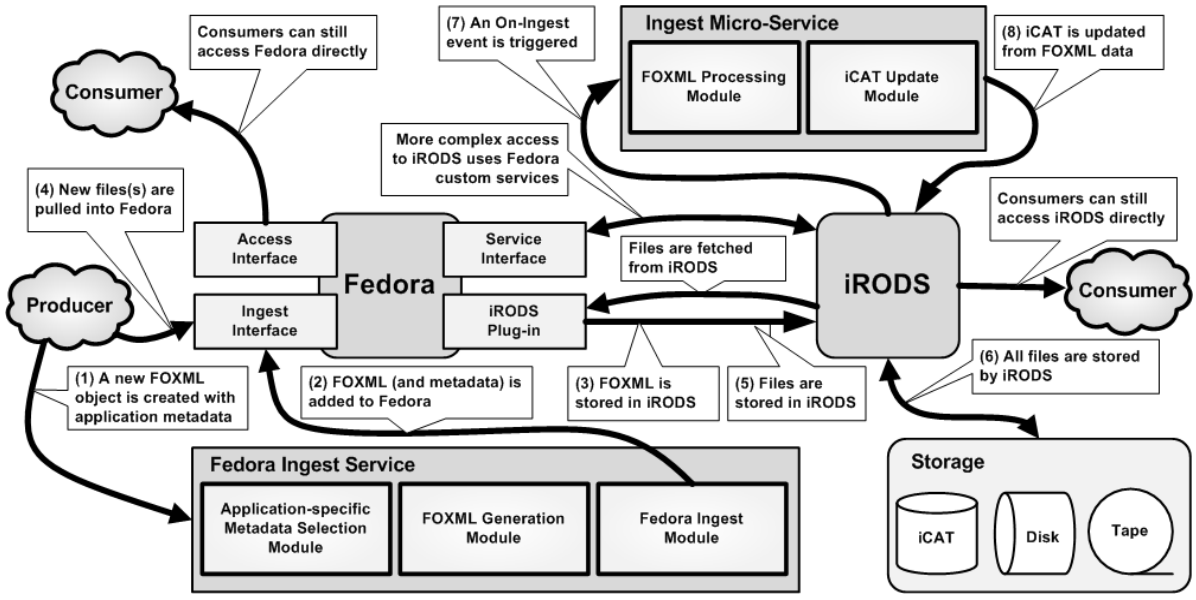


Figure 1: New Content Ingest via Fedora

metadata.

These interactions may trigger policy invocations. For example, Fedora may trigger policy invocation indirectly when interacting with a file (CRUD) or directly through a Fedora custom service. Conversely, iRODS' micro-services can call Fedora services to provide feedback in the system.

A more comprehensive "Concept of Operations" document will be prepared as part of the PoDRI project. The following set of questions is drawn from our current understanding of the operational scenarios:

- How will the collection structure be represented in the two products?
- How will Fedora be initialized for existing content in iRODS?
- How will Fedora be informed of content or metadata changes initiated directly in iRODS?
- How can content or metadata from Fedora be accessed by iRODS services?

4. ENABLING USE CASES

Five enabling use cases have been identified for the Fedora-iRODS integration. These use cases are:

1. New content ingest via Fedora
2. New content ingest via iRODS
3. Bulk registration from iRODS into Fedora
4. Update of content or metadata via Fedora
5. Update of content or metadata via iRODS

We introduce each of these use cases in this paper. While they do not by themselves represent policy management operations, they are prerequisites for enabling policy-driven operations and represent *demonstrations* of policy interoperability between repositories. The initial implementation work is focused on uses cases one and two together with the storage plug-in, a key enabler, described in Section 5.1.

4.1. New Content Ingest via Fedora

Current users of Fedora will want to continue ingesting into Fedora. Users are also likely to use Fedora features to add and relate rich metadata including policy, provenance and authenticity information. As shown in Figure 1, when new content is ingested into Fedora, it is able to capture the metadata needed for its operation. Digital content (or user-defined metadata) is either pulled in by Fedora or pushed to Fedora and stored in individual files. The file containing the FDO (FOXML) and the content files are subsequently stored in iRODS with no permanent storage directly managed by Fedora.

Selected metadata is collected by Fedora during the ingest process and stored in an internal system index implemented using a relational database. This database is used only to speed up access to content or bindings to services (formerly called disseminators). Optionally, metadata or notifications can be sent to index services such as semantic triplestores, search engines and OAI-PMH harvesters.

The Carolina Digital Repository (CDR) is using Solr/Lucene as the indexing and search engine for discovery of ingested content. Metadata is extracted during the ingest process from MODS and FOXML files.

Objects ingested via Fedora and stored in iRODS do not, by default, retain the logical tree structure of the original file system. Instead, CDR preserves the hierarchal structure of the file system via relations in the RDF triple store.

The arrangement of objects is achieved by creating FDOs representing the parent and child. The relationship

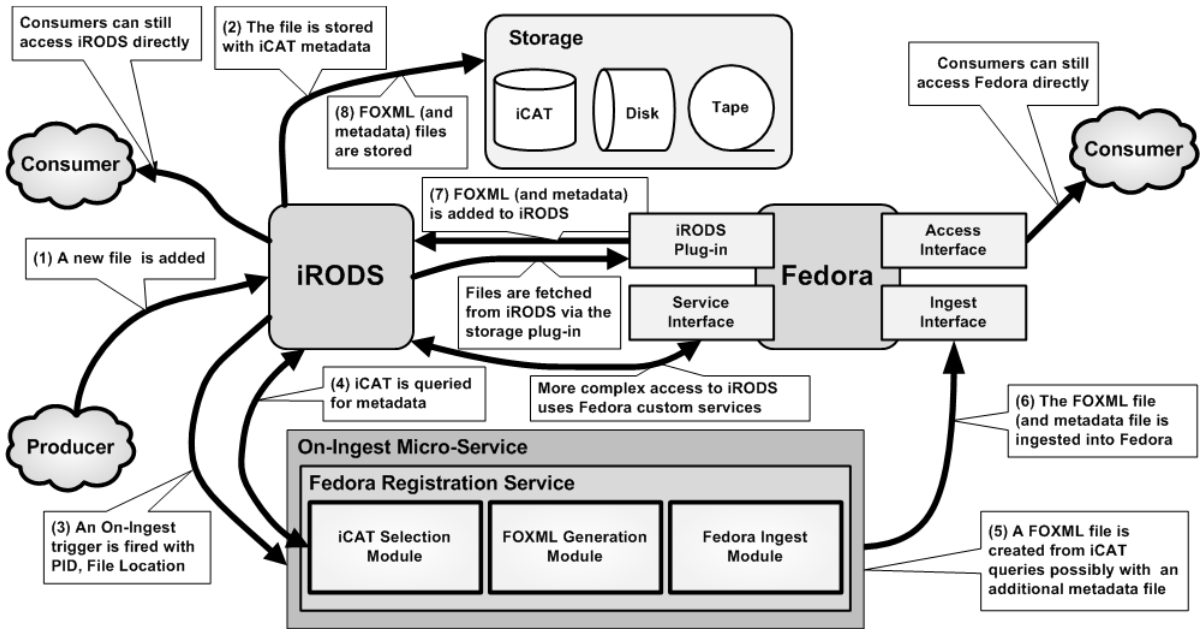


Figure 2: New Content Ingest via iRODS

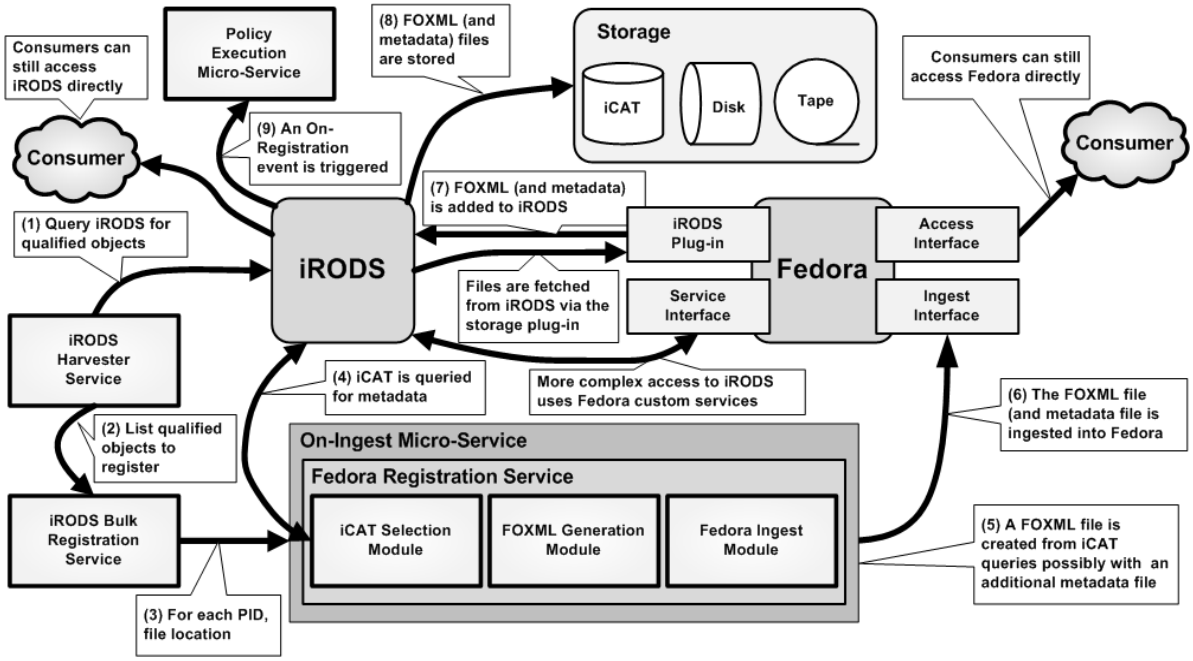


Figure 3. Bulk Registration into Fedora

is recorded in RDF (within the RELS-EXT Datastream) using the "isMemberOf" relation asserted in the child to the parent. The obverse relation "hasMember" is implied and could be stated explicitly in the parent. These two relations provide a way to build a hierarchical structure for all objects, collections and files. In Fedora, these relations form a "graph" and objects may participate in any number of graphs using other relations and, therefore, are not limited to a single hierarchy. Relationship information can be accessed by introspecting on the FDO or the relations can be indexed into a RDF triplestore [16] and queried by applications to extract a graph for navigating from parent to children as people usually do for a tree structure. Similar methods can be used to navigate any relationship graph.

How will the metadata in iRODS be updated in this use case? Two alternatives being considered are: (1) call a Fedora custom service to update the iCAT; (2) when the FOXML file is ingested, a monitoring rule can trigger an iRODS micro-service to introspect on the FDO to extract the metadata.

4.2. New Content Ingest via iRODS

Current iRODS users will likely want to continue to use iRODS directly to store data objects, particularly in research settings where direct access to storage is desired. The digital content (data object) is typically ingested into iRODS as a file operation. In iRODS, the hierarchical relation of a data object and its ancestors are encoded and described explicitly in its global object name. Two questions arise from this scenario. First, how will Fedora be notified of arrival of the new data object? Second, how will an analog component to its iRODS hierarchy be represented in Fedora?

As depicted in Figure 2, a utility is needed to register iRODS files into Fedora. A micro-service could call this utility when triggered by a monitoring rule on the storage operation which would create the FDO for the data object and ingest it into Fedora. The micro-service can be deployed as a rule under the iRODS rule event, 'acPostProcForPut'. Once this rule is activated in an iRODS server, the micro-service can be triggered after each new iRODS data object is created in a specified collection in the iRODS Content Store (see iRODS Storage Module). It will create pre-ingest FOXML for the new data object, querying the iCAT for additional metadata as needed. Within the FOXML, it will create a Datastream containing a reference to the location of the data object within iRODS. It will then ingest the FOXML using Fedora's API-M to create the FDO. This rule is activated once placed in the rule configuration file of an iRODS server. It will monitor all file activities in the iCAT catalog and will create an FDO for any newly created iRODS file.

When using iRODS for back-end storage, all FDOs and Datastreams are stored in iRODS as files in one of two collections: FOXML Object Store and iRODS Content Store. Therefore, users can directly access the files containing Fedora metadata through the iRODS interface. On the other hand, files stored in iRODS, whether for an FDO or a Datastream, have both an independent set of iRODS system metadata as well as a set of user-defined metadata. The system metadata contains important information for each replica of an iRODS file, including the file's location, storage type, audit trail, and associated iRODS rules. The two sets of metadata can be represented as external Datastreams in FOXML and generated dynamically when accessed using the Fedora-iRODS storage module.

As described above, Fedora uses RDF relations to describe the arrangement of objects. This requires the creation of FDOs representing each hierarchical level which has the advantage of enabling the participation of iRODS in the semantic network functionality provided by Fedora. Since iRODS can create a virtual hierarchy, it may not be desirable to instantiate corresponding FDOs. Users can create custom Datastreams as "finding aids"; the virtual hierarchy can then be encoded using RDF or any other desired format. Similar to iRODS, parent-child relationships can be modeled as path metadata and stored in the custom Datastream. An application or a Fedora custom service can be used to interpret the format of the Datastream to display the hierarchy [5].

One of the CDR's core constituencies are the special collections in our university libraries. These collections tend to have rich metadata associated with them and have usually undergone preliminary curation. The longer term goal of the repository is to harvest content directly from research-based iRODS data grids. Metadata quality and quantity is typically limited in these collections. Repository outreach and development is concerned not only with identifying and preserving "at risk" collections, but cultivating metadata collection and data curation proactively throughout the research lifecycle.

4.3. Bulk registration from iRODS into Fedora

Often we will be presented with existing collections in iRODS which we want to add to Fedora. How will these collections be registered into Fedora? It would be time consuming to require manual extraction, encapsulation (in an FDO) and storage of each data object.

As shown in Figure 3, a four step process is needed to automate this process: (1) identify the iRODS data objects to register; (2) iterate over each data object; (3) automatically collect metadata about each data object; (4) create the analogous FDO and ingest it via Fedora (possibly with additional FDOs to represent the hierarchy).

Bulk registration of a collection of iRODS files could be deployed and executed by a data curator through a

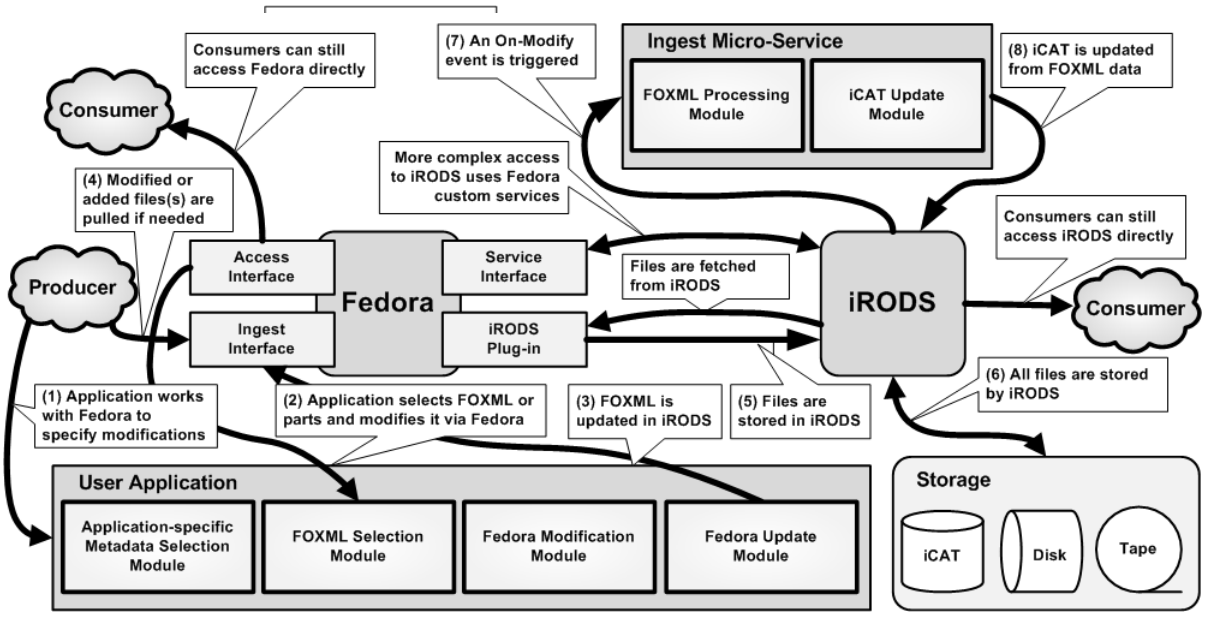


Figure 4. Update Content or Metadata via Fedora

single command irule, an iRODS command to send and execute a rule in an iRODS server. Often, such a rule is executed for a collection recursively. Registering multiple collections can be accomplished by through a batch script, which could query all iRODS files within a specified collection and create an FDO for each iRODS file. All FDOs could then be stored back into iRODS in the FOXML Object Store. Note that executing this process through iRODS facilitates inclusion of feature extraction services to automate metadata extraction. Also, note that services often can be reused in multiple use cases, reducing software development and deployment costs.

4.4. Update Content or Metadata via Fedora

Updates via Fedora use the same techniques as when iRODS is not present (see Figure 4). A user or application uses the Fedora APIs to update metadata, add new Datastreams, or new Datastream versions. Note that when a Datastream is versioned, it will result in a new file in iRODS.

However, iRODS must update system and user metadata in the iCAT by interpreting the updated FDO, extracting modified metadata, and updating the iCAT. This operation is triggered when the modified FDO is saved causing the execution of a micro-service to perform these tasks. The micro-service will then be able to query the iCAT to fetch data that will help and update the iCAT.

4.5. Update Content or Metadata via iRODS

An update via iRODS is similar to registering a new iRODS file into Fedora. As shown in Figure 5, however, the FDO already exists and must first be fetched, revised and updated in Fedora. The Fedora APIs are also capable of performing more fine-grained operations on the FDO for very common updates. For example, if a new file is added to iRODS, it may be registered in a new FDO (see *Ingest New Content via iRODS*). However, since Fedora supports a compound object model, the new file could be added as a new Datastream to an existing FDO. In this case, the addDatastream API method is preferred. Similar convenience functions exist for updating relations and certain metadata elements.

If a file is updated in iRODS, metadata (for example, the "last update" timestamp) for the Datastream in Fedora must also be updated. If a new version of a file is added to iRODS, the updateDatastream API method is used. This can only be used if the new version is represented by a new file in iRODS.

The update case puts a significant burden on the micro-service to determine the best approach to update Fedora. In particular, the micro-service must be informed of the FDO which correlates with the iRODS file. The iCAT will have to be extended to include the Fedora PID.

5. ADDITIONAL UTILITIES

We are implementing two key enabling utilities in addition to the functionality described above. First is an updated storage module as an iRODS-specific plug-in to replace Fedora's Low-level Store. Second is a harvester utility which can be used in both bulk registration and for disaster recovery.

5.1. iRODS Storage Module

We plan to store all files in iRODS. This will require an update of the existing iRODS-Fedora Storage Module or building a new one. Because this is a key enabler, work is concentrating on updating the existing iRODS plug-in replacing the Fedora Low-Level storage module. A new storage module is also being built, using Jargon and the Fedora Commons Akubra interface. Furthermore, a storage module is being developed by Aschenbrenner, based on the Merritt Storage System [2], in an iRODS community project. We are also closely following work in DuraCloud for integrating with cloud storage and service providers [13]. Selecting these candidates was based on a survey of available storage subsystems, finding a great proliferation of new approaches. Testing, however, eliminated all FUSE-based solutions as too unreliable except for the most lightweight usage. Building a new storage module, based on one or more of these existing technologies, would permit research on using it as a feedback path for policy operations including security policies.

When iRODS serves as a storage module for Fedora, the current design is to use two iRODS collections: (1) Fedora Digital Objects (FOXML) in the FOXML Object Store, and (2) content objects (Datastreams) in the iRODS Content Store. They are accessed through a single curator user account in iRODS. This makes it easier to distinguish between policies related to FDOs from those operating on content objects (Datastreams).

This approach, however, differs from the Fedora/Jargon/Merritt default of storing objects in folders based on a directory/file path and naming scheme. For the CDR and other existing implementations, a restructuring of objects into the segregated object store will be required. This will alter iRODS based failure recovery mechanisms and integrity audits.

5.2. iRODS Data Harvester for Fedora

The iRODS Data Harvester is an adaptive version of the Data Rebuilder in Fedora. It is used to re-build the object indices from the FOXML Object Store and iRODS Content Store. It does not create any new FOXML objects; rather, it surveys all the objects stored within the FOXML Object Store, verifies the Datastreams inside the iRODS Content Store, and creates the indices in the database used by the Fedora server. The iRODS Data Harvester also builds the

necessary RDF data to be stored in the RDF triplestore for the navigation of hierarchical structure.

6. POLICY FEDERATION AND MIGRATION

The iRODS rule engine provides the capability to apply rules on the data grid side to implement the policies. The Distributed Custodial Archival Preservation Environments (DCAPE) project [4] aims to work with a group of archivists to develop a set of rules to automate many of the administrative tasks associated with the management of archival repositories and validation of their trustworthiness. These DCAPE rules could be applied to different repositories based on the institution's policies. We plan to provide the functionality for users to manage the policies through the Fedora interface and be able to check what rules are in action.

Current implementations, even in data grid environments, depend on local enforcement of policies and typically do not consider the larger framework of uniform policy implementation across heterogeneous repositories. Though currently still in development, the ISO/NP 1636 standard [11] could present a model for identification of machine-actionable rules that can be expressed as policies. Stored as Fedora Service Definitions, the policies will have unique service deployment bindings for each data storage system. Our demonstration storage implementation is iRODS, but other storage environments may be supported by changing deployment mechanisms.

The CDR is developing a policy management framework based on a machine interpretable series of actions across repositories in a data grid. Implementation of new policy requires identification of machine-actionable components and mapping to specific, testable deployment mechanisms.

7. SUMMARY

In this paper, we introduced the Policy-Driven Repository Interoperability (PoDRI) project investigating interoperability mechanisms between repositories at the policy level. The rationale for using iRODS and Fedora to demonstrate key features of a distributed policy-driven management architecture was described. Four scenarios that will be demonstrated as part of the project were enumerated. We have identified five enabling use cases that are needed for the demonstration scenarios along with two key utilities planned for development. We also introduced work on policy federation and migration. PoDRI is an applied research project and its details will change as we develop a greater understanding of the methods for policy-driven interoperability.

8. ACKNOWLEDGEMENTS

This project is funded by IMLS grant LG-06-09-0184-09 as part of the 2009 National Leadership Grants NLG Library-Research and Demonstration, awarded to the University of North Carolina at Chapel Hill. Project Director is Richard Marciano. Collaborators at UNC / SILS include: Alex Chassanoff, Chien-Yi Hou, Reagan Moore, and Helen Tibbo. At UNC / Libraries: Steve Barr, Greg Jansen, Will Owen, and Dave Pcolar. At UNC / RENCI: Leesa Brieger. At UCSD: Bing Zhu. At DuraSpace and Cornell Information Sciences: Daniel Davis and Sandy Payette. Finally, at the University of Maryland iSchool: Bruce Ambacher.

9. REFERENCES

[1] Aschenbrenner, A., Zhu, B. "iRODS-Fedora Integration", http://www.irods.org/index.php/Fedora

[2] California Digital Library. "Merritt storage service", http://confluence.ucop.edu/display/curation/storage

[3] DART, University of Queensland. "Fedora-SRB Database integration module" http://www.itee.uq.edu.au/~eresearch/projects/dart/outcomes/FedoraDB.php

[4] DCAPE. "Distributed custodial archival preservation environments", an NHPRC-funded project, http://dcape.org

[5] DuraSpace. "The Content Model Architecture", http://fedora-commons.org/confluence/x/gABI

[6] DuraSpace. "The Fedora Digital Object Model", http://fedora-commons.org/confluence/x/dgBI

[7] DuraSpace."Fedora Repository 3.3 documentation", http://fedora-commons.org/confluence/x/AgAU

[8] DuraSpace. "PLEDGE project", http://fedora-commons.org/confluence/x/WSDS

[9] Fedora Commons, http://www.fedora-c ommons.org

[10] Hedges, M., Hasan, A., and Blanke, T. "Management and preservation of research data with iRODS", *Proceedings of the ACM first workshop on CyberInfrastructure:Information management in eScience*, Lisbon, Portugal, pp. 17-22, 2007. doi: http://doi.acm.org/10.1145/1317353.1317358

[11] International Organization for Standardization. "ISO/NP 16363: Audit and certification of trustworthy repositories", http://www.iso.org/iso/catalogue_detail.htm?csnumber=56510

[12] iRODS: Data grids, Digital Libraries, Persistent Archives, and Real-time Data Systems. http://www.irods.org

[13] The Library of Congress. "DuraCloud", http://www.digitalpreservation.gov/partners/duracloud/duracloud.html

[14] Moore, R., Rajasekar, A., Wan, M., and Schroeder, W. "Policy-based distributed data management systems", *The 4th International Conference on Open Repositories*, Atlanta, Georgia, May 19, 2009.

[15] NSF Cyberinfrastructure Council. "NSF's cyberinfrastructure vision for 21st century discovery", National Science Foundation, March 2007 http://www.nsf.gov/pubs/2007/nsf0728/index.jsp?org=NSF

[16] Wikipedia. "Triplestore", http://en.wikipedia.org/wiki/Triplestore

[17] Zhu, B., Marciano, R., and Moore, R. "Enabling Inter-Repository Access Management between iRODS and Fedora", *The 4th International Conference on Open Repositories*, Atlanta, Georgia, May 19, 2009

FULL PAPER

CHRONOPOLIS AND METAARCHIVE: PRESERVATION COOPERATION

David Minor
San Diego Supercomputer Center
UC San Diego
9500 Gilman Drive, MC 0505
La Jolla, CA 92093

Mark Phillips
UNT Libraries
University of North Texas
1155 Union Circle #305190
Denton, TX 76203

Matt Schultz
Educopia Institute
1230 Peachtree Street, Suite 1900
Atlanta, GA 30309

ABSTRACT

This paper will examine ongoing work between two major preservation systems, the Chronopolis Digital Preservation Program, [6] and the MetaArchive Cooperative. [13] In the past year, these two systems have begun work on bridging their technical underpinnings to create a more robust, reliable, long-lived preservation community for their users. The main emphasis of this work is moving data between a LOCKSS-based system (MetaArchive) and an iRODS-based one (Chronopolis). This work also involves several other emerging preservation micro-service tools and practices, and the expertise of the University of North Texas (UNT) Digital Library [21] in deploying them. The final result of this work is intended to be of three-fold benefit: 1) directly improving the services offered by Chronopolis and MetaArchive to their constituents; 2) offering specific technical findings which will be of benefit to other systems using LOCKSS and iRODS; and 3) contributing to the larger preservation community through the examination of organizational best practices for preservation system interactions.

1. BRIDGING METAARCHIVE AND CHRONOPOLIS

Large-scale digital preservation is a core technology need in many communities worldwide. The majority of information is now produced as digital files, rather than print output. To prevent the loss of significant cultural and scientific assets, active preservation systems must be put into place. This is not a theoretical threat: on a daily basis, data collections are lost for myriad reasons. The reasons for this range from the smallest and most mundane to the catastrophic, and they cannot be totally prevented–they are unavoidable in any large technology enterprise. Thus there is a core need to preserve data as rigorously as possible to make it live into the future.

Several projects and technologies are now focused on this need. Two of the most successful projects and their corresponding open source technologies are the Chronopolis Digital Preservation Program making use of the Integrated Rule-Oriented Data System (iRODS) [9] and the MetaArchive Cooperative making use of the Lots Of Copies Keeps Stuff Safe (LOCKSS) platform. [11]

1.1. Chronopolis and iRODS

Chronopolis is a digital preservation data grid framework developed by the San Diego Supercomputer Center (SDSC) at UC San Diego, the UC San Diego Libraries (UCSDL), and their partners at the National Center for Atmospheric Research (NCAR) in Colorado and the University of Maryland's Institute for Advanced Computer Studies (UMIACS).

A key goal of the Chronopolis framework is to provide cross-domain collection sharing for long-term preservation. Using existing high-speed educational and research networks and mass-scale storage infrastructure investments, the partnership is designed to leverage the data storage capabilities at SDSC, NCAR and UMIACS to provide a preservation data grid that emphasizes heterogeneous and highly redundant data storage systems.

Specifically, the current partnership calls for each Chronopolis member to operate a grid node containing at least 50 TB of storage capacity for digital collections related to the Library of Congress' National Digital Information Infrastructure and Preservation Program (NDIIPP). [14] For reference, just one terabyte of information would use up all the paper made from about 50,000 trees. The Chronopolis methodology employs a minimum of three geographically distributed copies of the data collections, while enabling curatorial audit reporting and access for preservation clients. The original underlying technology for managing data within Chronopolis has been the Storage Resource Broker, [20] a preservation middleware software package that allows for robust management of data. The partnership is also developing best practices for the worldwide preservation community for data packaging and transmission among heterogeneous digital archive systems.

Chronopolis has concentrated on building a wide range of content that is not tied to a single community. Currently there are four significant collections housed in Chronopolis. These include:

- A complete copy of the data collection from The Inter-university Consortium for Political and Social Research (ICPSR), based at the University of Michigan. Established in 1962, ICPSR is the world's largest archive of digital social science data. [10]
- Data from The North Carolina Geospatial Data Archiving Project, a joint project of the North Carolina State University Libraries and the North Carolina Center for Geographic Information and Analysis. It is focused on collection and preservation of digital geospatial data resources from state and local government agencies in North Carolina. [15]
- Scripps Institution of Oceanography at UC San Diego (SIO) has one of the largest academic research fleets in the world, with four research vessels and the research platform FLIP. Since 1907, Scripps oceanographic vessels have played a critical role in the exploration of our planet, conducting important research in all the world's oceans. SIO is providing data from several decades of data from its cruises. [18]
- The California Digital Library (CDL) is providing content from its "Web-at-Risk" collections. Web-at-Risk is a multi-year effort led by CDL to develop tools that enable librarians and archivists to capture, curate, preserve, and provide access to web-based government and political information. The primary focus of the collection is state and local government information, but may include web documents from federal and international government as well as non-profit sources. [5]

Chronopolis is currently transitioning from the use of SRB to iRODS. One of the hallmarks of iRODS is its rule-based architecture. On top of an advanced preservation environment, this rule-based architecture allows iRODS administrators to create a customized environment that follows designated rules and triggers specific actions based on certain events.

The rule-based process has three layers. The most granular layer is a system of micro-services. In the iRODS context, micro-services are functions that have been written to accomplish a certain task. A large set of micro-services ships with the default iRODS installation, but additional ones can be written by iRODS systems administrators as needed in their particular environment.

Micro-services can be chained together to form longer processes called actions. Actions are macro-level tasks that typically call on multiple micro-services. Actions are called or started based on predefined rules. These rules are tasks that the iRODS system needs to perform when certain conditions are met. The iRODS system has a built-in rule-engine that then interprets rules and calls the underlying actions (and hence the micro-services) when appropriate.

An example of an iRODS rule: when a new file of type x is added to the system, rename it adding a timestamp to its filename and copy it to another location. The rule in this case is calling two actions (renaming process and copying process). Each of these actions consists of multiple micro-services (which do the actual underlying work to make changes to the file and file system).

1.2. The MetaArchive Cooperative and LOCKSS

Originally created as an initiative of the US National Digital Information Infrastructure and Preservation Program (NDIIPP), the MetaArchive Cooperative is a distributed, nonprofit-based alliance of university libraries, archives, and research centers. The Cooperative's purpose is to support, promote, and extend distributed digital preservation practices. Since 2004, the MetaArchive Cooperative has provided community-owned and community-governed digital preservation activities through running a distributed preservation network that is based on the LOCKSS software.

To preserve digital assets, the MetaArchive Cooperative uses a systemic, forward-looking technological approach called distributed digital preservation. The member institutions identify collections that they want to preserve. They then ready these collections for preservation, creating Submission Information Packages (SIPs). Using a technical framework that is based on the LOCKSS software, these collections are then ingested into a geographically distributed network where they are stored on secure file servers in multiple locations that are housed by the member institutions. These servers do not merely back up the materials. Rather, they provide a dynamic means of constantly monitoring content via the LOCKSS software and its use of ongoing cryptographic SHA-1 hashes to compare the copies, determine if any have degraded in any way, and then provide repairs whenever necessary. Such redundancy and monitoring activities minimize the risk that information might be lost due to human error, technology failure, or natural disaster.

The Cooperative currently is comprised of seventeen member institutions that preserve their digital collections in a 254 TB network that is distributed internationally at thirteen distinct sites. The network grows both in content and in size as new members join the Cooperative. Its membership doubled in 2009, and it is expected to double again in 2010.

The Cooperative's mission is twofold: 1) providing distributed digital preservation services for its member organizations and 2) having an impact on the broader cultural memory field through modeling the use of open source technology and community-based infrastructures to accomplish digital preservation in ways that can be replicated by other groups.

To these ends, the Cooperative maintains transparency in its operations and makes available to other groups that seek to implement preservation solutions all of its administrative and technical developments. In this way, the Cooperative has fostered the formation and growth of other Private LOCKSS Networks (PLNs), [17] including the Persistent Digital Archives and Library System (PeDALs) initiative, [16] the Alabama Digital Preservation Network (ADPNet), [1] and the Data-PASS network, [7] run by the Interuniversity Consortium for Political and Social Research (ICPSR) at the University of Michigan. [10] It also recently published a book, *A Guide to Distributed Digital Preservation*, [19] which is intended to help other groups form and run their own distributed digital preservation networks.

1.3. UNT and CDL Micro-Services

Beyond these two successful projects and technologies, yet another new suite of preservation and curation tools that are proving integral to this work is being hosted at the California Digital Library (CDL), named the Curation Micro-Services. [4] According to the University of California's Curation Center, "micro-services are an approach to digital curation based on devolving curation function into a set of independent, but interoperable, services that embody curation values and strategies." These small and self-contained services span the range between providing persistent URLs, unique identifiers, file system conventions, fixity checking, format migration and file transfer specifications, among many others.

The University of North Texas (UNT) was chosen as the key bridge technology partner in this interoperability work because they have demonstrated the great potential for putting these and several other micro-services into unbundled and modular use on behalf of transporting and managing digital objects and collections. UNT has constructed a robust and loosely integrated set of in-house archiving infrastructures to manage their own digital collections, including a delivery system (Aubrey) and a repository structure (CODA). The underlying file system organization of digital objects is tied to a UNT-specific data modeling process that relies on locally developed scripts and CDL micro-services to generate and define all master, derivative, related objects, metadata, and other information that may be tied to a single digital object in order to effect timely archival management and access retrieval. This archival repository solution has been designed in a highly open source fashion and relies on loosely bundled specifications to ensure on-going flexibility and scalability.

1.4. Scope of Work

Each of these sets of technologies has strengths and weaknesses, but one action that would improve them all is the ability to transfer preserved objects between systems based on these technologies. Making this possible would offer a more robust suite of interoperable tools and allow preservation systems to leverage the power of each technology in a modular fashion. It would also enable practitioners using these systems to take advantage of tools and services created by any of these technologies.

The focus of this paper is to examine one instantiation of this transfer process, using already existent collections and trustworthy processes. The work that has already been done, and which will refined in the coming year, is based on daily use of the MetaArchive LOCKSS-based and the Chronopolis iRODS-based systems, and making use of BagIt, [2] a CDL micro-services based component, and other modular approaches, to efficiently facilitate a transfer. The collections being utilized are real, and the processes represent actual tasks.

The work being described has been made possible thanks to a grant provided by the National Historical Publications and Records Commission (NHPRC). The work put forth has been to successfully identify the necessary technologies and workflows needed to efficiently retrieve and package a complete collection from a LOCKSS file system, through the use of custom developed scripts and the BagIt specification, and maintain its archival unit integrity both structurally and at the file object level while transferring into a non-LOCKSS based environment for the purposes of providing a succession pathway. Fixity checking is required on the collection prior to initial retrieval from the LOCKSS file system, and validation is required both prior to packaging, and upon un-packaging on its destination directory registered in the iRODS storage environment managed by Chronopolis. Additional effort will be made to explore the packaging and transfer requirements for the MetaArchive's data management tool, known as the Conspectus, [12] as well as its associated collection level metadata.

2. STAGE ONE: COMPLETE

Chronopolis and MetaArchive have completed an initial round of testing the process of sharing data between their systems. This first round focused on transferring data from the MetaArchive LOCKSS-based system into Chronopolis' SRB-based system. This was done using two different transfer approaches.

2.1. BagIt-Based Transfers

First, the BagIt tool was used as a simple proof-of-concept on behalf of four test collections of data of approximately 200MB. BagIt is a simple packaging specification that incorporates a human-readable manifest file. This file lists the digital objects in the package as well as their checksums and serves as an authoritative inventory list. Between July 15 and August 11, 2009 system administrators from the MetaArchive

and Chronopolis worked together to transfer archival units (AUs), measuring in the 100s of MBs, from the MetaArchive network into the Storage Resource Broker using what are known as BagIt files. The BagIt file specification allows for a regular bag and a "holey bag." A regular bag bundles up the actual data in a file directory, while a holey bag uses URLs that point to the data and performs an extraction.

These BagIt transfers (four Bag files in all) were of a small enough size to facilitate unsophisticated http "get" requests and even an email-based transfer to get the AUs into Chronopolis' SRB-configured storage environment. Upon completion the administrators verified the successful transfer of these individual Bags into SRB, ran checksum-based comparisons on the Bag content, and registered the content into their MCAT database (which captures and holds metadata that can be exported later for data provider purposes).

2.2. SRB Client-Based Transfers

Following this initial test with BagIt, An additional transfer was performed using a combination of custom-written and SRB-based client scripts as well as BagIt. Chronopolis staff first provided a script that gathered MetaArchive content into a "holey" bag. The SRB-specific scripts that function as Unix commands were then used to facilitate a "put" of those files to the MetaArchive's directory in SRB.

The MetaArchive system administrator was then required to download and install the client and set-up two specific files: an Environment file and an Authentication file:

- The Environment file sets up user credentials for the home directory on the assigned SRB storage environment. This is the location to which a Bag can be sent and unpackaged for quality control.
- The Authentication file stores a password to manage access to this environment.

2.3. Lessons Learned from Initial Transfers

Several lessons were learned from these initial processes, which are informing next steps for the project.

- MetaArchive staff had to iteratively work through several authentication and registration issues when setting up appropriate working and home directories in the designated SRB instance.
- During holey BagIt tests there were minor extraction issues related to LOCKSS. LOCKSS puts a '#' character in the directory structure that it creates. The '#' is treated as an html anchor, and this causes problems during a web transfer. To surmount this it was necessary to URL encode the '#' and turn each one into a '%23'.
- MetaArchive AUs and/or complete collections must be taken out of active preservation mode and be rendered static before being placed into Bags and transferred to Chronopolis, otherwise the LOCKSS re-crawling and polling/voting process(es) will interfere with their packaging.

Also, based on these lessons, several areas of refinement were designated for the next stage of work:

- The need to measure transfer rates as data flows between the systems, especially to help determine if one method is more efficient or provides better service.
- Usability comparisons between use of an SRB (now iRODS) client transfer and that of a manual send/get of BagIt files through standard web channels.
- Transferring collections in excess of 1TB to achieve large-scale efficiency.

3. STAGE TWO: CURRENT PROCESSES

Based on what was learned in these initial steps, the current processes were begun, with several guiding principles in mind. The first of which regarded the feasibility of transferring MetaArchive collections on a larger scale to Chronopolis's data grid environment (now running on iRODS) it was decided to do so at a larger AU or collection level. SRB and iRODS, using BagIt, can handle ingests of content in the multiple TB range.

From an ease of packaging and transfer perspective, it was initially encouraged to use a true bridge server (non-LOCKSS based), so that content can be migrated in a static condition via the LOCKSS content serving feature or through a WARC, ARC or ZIP extraction. Bags can then be generated and sent from this bridge server via an installed iRODS client. This avoids interference from the routine LOCKSS operations on a cache that may impede a transfer.

Based on these recommendations, beginning in April 2010, efforts were begun to improve the transfer of MetaArchive collections through addressing the items listed above. This phase of work is relying on Chronopolis' new iRODS configuration, but still makes use of BagIt as the primary transfer mechanism.

3.1. Larger Collection

For this phase a new, larger MetaArchive collection has been designated. The Folger Shakespeare Library has agreed (through an MOU) to permit the use of a copy of their 1.5TB collection currently being preserved in the MetaArchive network. A MetaArchive-LOCKSS cache located at the University of North Texas (UNT) will harvest this collection. A developer from UNT will prepare the Folger digital collections for transfer to Chronopolis, manage this transfer with tests for content integrity and authenticity, and address the above lessons learned and areas for refinement. Staff at Chronopolis will coordinate with UNT's staff to receive, validate, and preserve the Folger content, and also facilitate with addressing the above "objectives."

3.2. Stage Two Summary of Work

The following tasks are slated for completion in this current work process:

- UNT will bring up its MetaArchive cache in consultation with MetaArchive staff;
- UNT will harvest the Folger Shakespeare Library collection, and validate its integrity through the LOCKSS voting/polling measures;
- UNT will collaborate with Chronopolis to transfer the Folger collections from MetaArchive's LOCKSS-based network to Chronopolis's iRODS-based preservation service and back again.

This work will serve is serving as a proof-of-concept that the MetaArchive network may use Chronopolis's iRODS-based preservation service as an exit strategy in the event that either MetaArchive or LOCKSS becomes unsustainable in the future.

3.3. Stage Two Summary of Progress

As of July 2010, the following measures have been accomplished ahead of enacting a full-scale second transfer of MetaArchive collection content into the Chronopolis environment:

- UNT configured a 50TB server on-site as a MetaArchive-LOCKSS cache in order to host the 1.3TB Folger collection;
- UNT coordinated with MetaArchive member GA Tech to proxy export the full Folger collection and metadata onto its MetaArchive-LOCKSS cache;
- UNT's cache participated a full round of LOCKSS-driven file voting/polling validation and ensured 100% integrity of Folger collection content;
- UNT developed and tested a custom script that exploits the in-built LOCKSS content serving features and standard HTTP protocols, and relies upon open source micro-services such as httplib2, Beautiful Soup, and other Python libraries to retrieve and validate the Folger files, and package each archival unit according to the "holey" BagIt specification;
- Chronopolis has provided and configured an iRODS client tool for UNT and registered a storage resource within their San Diego SuperComputer Center data node environment;
- Preliminary transfer rates were tested on a 6GB archival unit subset of Folger collection content and it was determined that the entire 1.3 TB could be transferred over the course of a 48 hour period;
- UNT, Chronopolis and MetaArchive staff began evaluating requirements for ensuring that the Conspectus data management tool and its associated collection level metadata could be exported into the Chronopolis environment.

3.4. Additional Work

In addition, discussion has begun between the groups toward developing strategies for how data can be transferred out of Chronopolis' iRODS environment and into MetaArchive's LOCKSS based storage. So far this has involved a preliminary examination of which iRODS rules may be necessary to stage the sharing of data between an iRODS and a LOCKSS environment. This analysis will continue to involve developing a better understanding of the differences in file systems, file naming conventions, directory structures, and file movements within the systems. Each of these differences will likely impact the kinds of micro-services, actions and rules that are needed. We anticipate that some of the available default micro-services will be part of the process, but that significant custom work will also be needed. In addition the project will need to keep track of which metadata is specific to each of the systems and which might need to be added or modified based on the iRODS actions.

4. FUTURE WORK

The ability for different digital preservation solutions to interoperate is necessary to reach the goal of long-term preservation of digital resources. The interchange of content between two repositories such as the MetaArchive Cooperative and Chronopolis stands as a use case for future work in the area of interoperability of digital preservation system for sustainability purposes. The work that will be accomplished in the next year lays the groundwork for future detailed, deep work to share preservation objects among diverse systems. Several specific next steps in this area include having a better understanding of the optimal granularity of units being passed between the two systems, identifying any needed data management implementations for ensuring best practices for administrative, technical, structural and preservation metadata, as well as the requirements that end users may have for retrieving archived content from these preservation networks and re-creating collections at their local institutions.

5. REFERENCES

[1] Alabama Digital Preservation Network (ADPNet). Available at: http://www.adpn.org/

[2] BagIt File Packaging Format. Available at: http://www.cdlib.org/inside/diglib/bagit/bagitspec.html

[3] Beautiful Soup. Available at: http://www.crummy.com/software/BeautifulSoup/

[4] California Digital Library: Curation Micro-Services. Available at: http://www.cdlib.org/services/uc3/curation/

[5] California Digital Library: "Web-at-Risk". Available at: http://www.cdlib.org/services/uc3/partners/webatrisk.html

[6] Chronopolis: Preserving Our Digital Heritage. Available at: http://chronopolis.sdsc.edu/

[7] Data Preservation Alliance for the Social Sciences (Data-PASS). Available at: http://www.icpsr.umich.edu/icpsrweb/DATAPASS/

[8] Httplib2. Available at: http://code.google.com/p/httplib2/

[9] Integrated Rule-Oriented Data System (iRODS). Available at: http://www.irods.org/

[10] Inter-university Consortium for Political and Social Research (ICPSR). Available at: http://www.icpsr.umich.edu/icpsrweb/ICPSR/

[11] Lots of Copies Keep Stuff Safe (LOCKSS). Available at: http://www.lockss.org/lockss/Home

[12] MetaArchive Conspectus Tool. Available at: http://conspectus.metaarchive.org/archives/list

[13] MetaArchive Cooperative. Available at: http://metaarchive.org/

[14] National Digital Information Infrastructure and Preservation Program (NDIIPP). Available at: http://www.digitalpreservation.gov/library/

[15] North Carolina Geospatial Data Archiving Project. Available at: http://www.lib.ncsu.edu/ncgdap/

[16] Persistent Digital Archives and Library System (PeDALS). Available at: http://pedalspreservation.org/

[17] Private LOCKSS Networks. Available at: http://lockss.stanford.edu/lockss/Private_LOCKSS_Networks

[18] Scripps Institution of Oceanography at UC San Diego (SIO). Available at: http://scripps.ucsd.edu/

[19] Skinner, Katherine and Matt Schultz Eds. *A Guide to Distributed Digital Preservation*, Educopia Institute, Atlanta, GA, 2010. Available at: http://www.metaarchive.org/GDDP

[20] Storage Resource Broker (SRB). Available at: http://www.sdsc.edu/srb/index.php/Main_Page

[21] University of North Texas Digital Library. Available at: http://digital.library.unt.edu/

Session 8b: Preserving Web Data

SHORT PAPER

PRESERVING VISUAL APPEARANCE OF E-GOVERNMENT WEB FORMS USING METADATA DRIVEN IMITATION

Jörgen Nilsson
Luleå University of Technology
Dept. of Business Administration and Social Sciences

ABSTRACT

This paper summarizes work done in a PhD study on metadata driven imitation for preservation of visual appearance of web forms and/or receipts used in eGovernment services. The research done suggests that metadata, and e.g. a background image, can be used to describe the visual appearance of documents, and that this also facilitates having the data separated from the visual appearance. This separation provides the ability to present the material to the users in different ways, depending on their needs and requirements, while retaining the ability to present the object in its original look. The original look is seen as the most versatile way of presenting the material, giving the most fruitful base of interpretation and understanding, but if the users where familiar with the material, they liked the ability to have the material presented in simplified ways, where many of the sometimes "distracting" visual attributes where removed. In general, preserving the visual appearance and keeping the data separated from the form, was seen as useful and beneficial to both the users and the preservation professionals. As always in digital preservation contexts, documentation of this process and the relation between the metadata describing the visual appearance and the data of the document, is of high importance.

1. INTRODUCTION

In recent years, the ongoing eGovernment proliferation of public administration has taken great steps toward availability and sophistication. The eGovernment Benchmark Survey 2009 [4] shows that the overall level of full online availability of 20 basic services in the EU27+ has risen from 59% to 71% between 2007 and 2009. The sophistication of the services has risen from 76% to 83% in the same period of time [4]. These are average numbers for the EU27+, some countries have achieved 100%, and yet some are over 90% in both categories.

This increase reasonably means that there will be an increase in the number of digitally born documents/records in need of preservation. Some organisations might also need/want/be obligated to preserve the visual appearance of these documents, and maybe also the appearance of the services, in order to fulfil expectations and demands from their designated community. There can be numerous reasons to preserve the visual appearance of digitally born documents, and some of them can be found in reasoning around the concept of *information*.

The concept of information has in this work been influenced by the *infological equation* (1) which states that *information* (I) is the result of an interpretation process (i) that acts upon data (D) involving the parameters of pre-knowledge (S) and time (t) [12].

$$I = i(D, S, t) \quad (1)$$

One important implication of the infological equation is that data does not contain information but at best can represent information to those who have the required pre-knowledge [12]. In addition to this data also acts as constraining affordances where data allows some constructs of information and impede others and that these constructions might differ between individuals [5]. Since humans interpret data, and occasionally with different results, as much of the original data should be available in order to give good basis for similar interpretations by different individuals. Part of this original data can exits in the form of visual attributes, such as colour, italics, tables and other layout properties.

This has lead to an interest in preserving "looks" of web resources, especially those created in eGovernment services.

2. PRESERVATION OF WEB

There are (at least) two approaches to web preservation. One approach consists of gathering the web-site(s) with a crawler accessing the web as a client and thereby fetching the web from a user perspective by following links. A typical drawback with crawling would be that it does not fetch documents that you as a user would need to fill in a form to fetch (i.e. deep web), for example by searching in an article database [9].

Another approach to gathering the web would be to keep the server side of the web intact, meaning that the web site still could be accessible in it's original way, as long as the ability to run the entire server side, including e.g. databases, still exists [9]. This could be facilitated by the use of emulation or migration depending on the requirements of the organisation. The emulation approach has for quite some time been proposed as *the solution* [15], but as pointed out both migration and emulation is not yet mature enough for large scale preservation scenarios, although it usually is better than doing nothing [9].

2.1. Significant properties

This paper assumes that the visual appearance or physical structure of a digital object (e.g. a web form) is considered to be a significant property of the object. This may of course differ from case to case as with all significant properties [2], and is certainly not true in all preservation of digital objects. Significant properties are "those components of a digital object deemed necessary for its long-term preservation" [2]. This is a quite common view of significant properties [6],[11],[16], held on a generic level since it is hard to be specific about significant properties in writing unless you actually consider one particular object or group of objects.

One way to handle significant properties have been addressed in work with the Underlying Abstract Form (UAF) [8]. The UAF holds "all the significant properties of the data, and is independent of the medium upon which the data is written" [8], and although not mentioning metadata or physical structure explicitly, they do suggest utilizing the representation information container in the OAIS model to hold the UAF, which implies using metadata, even though it could be as simple as referring to a viewer application for the data object e.g. Acrobat Reader for a pdf-file. The UAF prefers to have the representation information pointing out the original software used to access the data object, and that this software also should have been preserved. And although "enabling meaningful access to the preserved object includes such processes as recreating the experience of viewing (or even interacting with) the original" [8], the author of this paper however prefer to focus on the viewing part, using an abstraction of the original objects presentation, described with the aid of metadata and e.g. screen dumps, since the original software could mean that you, for good or bad, preserve the system instead of the information [1], meaning that the users in the future would need to know how to use old software in order to access the information. The approach suggested below instead allows for several different ways of presenting the material to the user, depending on their needs and wants.

2.2. Preserving physical structure of deep web documents

Although deep web can contain lot of different types of digital objects, a respectable amount of the objects created in eGovernment context would likely be of a textual character related to filling out web based forms. Some of the objects may be e.g. pdf-files submitted as attachments to a web-form, but still – the actual web form would also have some content filled in and saved, most likely, in a database. This implies that we already here have a separation of the physical structure and the data, and when they are combined together again we get the digital object in its original shape [14] or performance [7].

The separation of physical structure and data makes it possible to treat the respective components according to their preservation needs. However, if the intention is that the original shape of the object should be possible to present again to the designated community, you do need to retain the ability to combine them together again in the future, regardless of what preservation actions they have been subjected to.

One way of addressing this re-presentation is to use *metadata driven imitation* [13] where the physical structure is described by a combination of layout metadata and e.g. backdrop images making out the main part of the layout. One could argue that this poses problems regarding the integrity of the document, but as pointed out in the InterPARES project, “a record has integrity when it is complete and uncorrupted in all its essential respects” [10] meaning that the record does not need to be exactly the same as when it was created, as long as the message it communicates remains is unaltered [10].

The type of metadata driven imitation that is mentioned here is most suitable for documents that appear in large numbers with similar physical structure, in other words, typical forms filled out in eGovernment contexts. Bearing in mind that these types of objects usually are not available to web crawling, these deep web objects need to be collected in some other way.

By describing the layout with metadata and background images, the data can then be linked (again, with metadata) to the layout in order to be presented upon request as a "whole". This also facilitates making other sorts of presentations to fulfill requests from different user communities, where some may only need

e.g. a particular data field from thousands of forms, while others are more interested in a complete form with its visual appearance as intact as possible. These kinds of diverse user communities are likely customers of e.g. large national institutions such as national archives or national libraries where the *general public* is the *designated community*.

3. OPINIONS ON METADATA DRIVEN IMITATION

Studies done on what potential users and preservation professionals think about the approach with metadata driven imitation [14] shows some interesting results that are presented below.

Most users preferred to have the data presented in a simplified form, where some visual attributes where removed (e.g. background colours and logotypes) while the layout in general (i.e. the physical relation between the data elements) remained intact. It should however be noted that the respondents said that the original look would give the best possibilities for interpretation, depending on the users familiarity with the material. The preservation professionals did prefer the original look, for the same reason as the users; it provides the best basis for a "correct" interpretation. This can be put in relation to the constraining affordances of data, which both facilitates and limits the interpretations possible.

Both the professionals and the users liked the ability to present the material in different ways, depending on the needs of the user. Some would for example only need the data, and cared less for the look of the document for their own purposes, but they also recognized the importance of retaining the ability to represent the document in its original form. The flexibility in presentation is facilitated by the separation of data from physical structure, and as pointed out by the preservation professionals, this separation also facilitates the ability to handle the data and the physical structure in different ways from a preservation perspective.

The separation mentioned above was recognized as a good feature from a slightly different perspective as well. The ability to only fetch data from a document, mean that it is quite easy to request the same kind of data from a large number of documents, for e.g. statistical purposes, instead of having to extract the data from an actual document, perhaps in an entirely manual way (i.e. actually *reading* the documents). So, although original look was regarded as important in general, the ability to choose from several different ways of presenting the data was seen as valuable. The objects used as demonstrators did not represent the *feel* of the documents, and the users did not see feel as that important on document level, though it certainly can be important at a system level, if that is what you are preserving.

Questions were also posed about the relation between original look and trust. Though the users said that the most trustworthy representation was the original look, they also realized that this might be a false sense of trust. They also pointed at that the trust mainly lies in that they trust the organization that manages the objects, and that they thereby probably would not question a document coming from them that much, in case they did not actually see something that they know is wrong. The preservation professionals, and some of the users, where careful to point out that you must have documentation about the processes concerning the material, for example about how the metadata descriptions of visual attributes are constructed, and used, so that the knowledge about this does not disappear over time.

To sum it up it;

- keep data and physical structure separated for usefulness and flexibility
- find a "middle way" of representing physical structure of the document type in question (e.g. by using a background image for capturing some of the physical structure)
- document everything that the object is subjected to

It can in general be summed up as, yes visual appearance of web forms in eGovernment context is important to preserve, since it both provides more context and acts as constraining affordances and thereby may facilitate better interpretation of the data into the intended information. However, fixing the data to a physical structure may impair the ability to mass process it, and therefore a separation of data from its physical structure would be beneficial. One way of addressing these issues can be by using *metadata driven imitation.*

4. REFERENCES

[1] Bearman, D. "Reality and chimeras in the preservation of electronic records." *D-Lib Magazine 5(4).* 1999. Retrieved 2004-03-13 from http://www.dlib.org/dlib/april99/bearman/04bearman.html

[2] Cedars Project. *Cedars Guide to Digital Collection Management*. 2002. Retrieved 2008-04-08 from http://www.leeds.co.uk/cedars/guideto/collmanagement/guidetocolman.pdf

[3] Dollar, C.M. *Authentic electronic records: Strategies for long-term access*. Cohasset Associates Inc., Chicago, IL, 2000.

[4] European Commission. *Smarter, Faster, Better eGovernment*, Brussels, Belgium, 2009.

[5] Floridi, L. "Semantic Conceptions of Information". *The Stanford Encyclopedia of Philosophy* (Fall 2008 Edition), Edward N. Zalta (ed.). 2008.

http://plato.stanford.edu/archives/fall2008/entries/information-semantic/

[6] Hedstrom, M. & Lee, C. "Significant properties of digital objects: definitions, applications, implications", *Proceedings of the DLM-forum 2002 Access and preservation of electronic information: Best practices*. pp. 218-223. European Communities. 2002. Retrieved 2006-05-28 from http://ec.europa.eu/comm/secretariat_general/edoc_management/dlm_forum/doc/dlm-proceed2002.pdf

[7] Heslop, H. Davis S. Wilson, A. *An Approach to the Preservation of Digital Records*, National Archives of Australia, Canberra. 2002 Retrieved 2004-03-03 from http://www.naa.gov.au/recordkeeping/er/digital_preservation/Green_Paper.pdf

[8] Holdsworth, D. & Sergeant, D.M. *A Blueprint for Representation Information in the OAIS Model*, The Cedars Project, 2000. Retrieved 2004-03-15 from http://esdis-it.gsfc.nasa.gov/MSST/conf2000/PAPERS/D02PA.PDF

[9] International Internet Preservation Consortium. *Long-term Preservation of Web Archives – Experimenting with Emulation and Migration Technologies*. International Internet Preservation Consortium, 2009. Retrieved 2010-04-25 from http://netpreserve.org/publications/NLA_2009_IIPC_Report.pdf

[10] InterPARES. *The Long-term Preservation of Authentic Electronic Records: Findings of the InterPARES project. Appendix 2 – page 2*. InterPARES Project, 2002. Retrieved 2007-05-10 from http://interpares.org/book/interpares_book_k_app02.pdf

[11] Knight, G. *Framework for the definition of significant properties.* InSPECT project. 2008. Retrieved 2008-04-08 from http://www.significantproperties.org.uk/documents/wp33-propertiesreport-v1.pdf

[12] Langefors, B. *Essays on infology: summing up and planning for the future*. Studentlitteratur, Lund, 1995.

[13] Nilsson, J. & Hägerfors, A. "Metadata Driven Presentation of Digital Documents/Records", *Constructing and Sharing Memory: Community Informatics, Identity and Empowerment*. Stillman, L. & Johanson, G. (ed.), Cambridge Scholars Publishing, 2007.

[14] Nilsson, J. *Preserving Useful Digital Objects for the Future*. Luleå University of Technology, Luleå, Sweden, 2008

[15] Rothenberg, J. & Bikson, T. *Carrying Authentic, Understandable, and Usable Digital Records Through Time*. The Dutch National Archives and Ministry of Interior, 1999. Retrieved 2008-04-27 from http://www.digitaleduurzaamheid.nl/bibliotheek/docs/final-report_4.pdf

[16] Wilson, A. *Significant properties report.* InSPECT project. 2007. Retrieved 2008-04-08 from http://www.significantproperties.org.uk/documents/wp22_significant_properties.pdf

UROBE: A PROTOTYPE FOR WIKI PRESERVATION

Niko Popitsch **Robert Mosser** **Wolfgang Philipp**

University of Vienna
Faculty of Computer Science

ABSTRACT

More and more information that is considered for digital long-term preservation is generated by Web 2.0 applications like wikis, blogs or social networking tools. However, there is little support for the preservation of these data today. Currently they are preserved like regular Web sites without taking the flexible, lightweight and mostly graph-based data models of the underlying Web 2.0 applications into consideration. By this, valuable information about the relations within these data and about links to other data is lost. Furthermore, information about the internal structure of the data, e.g., expressed by wiki markup languages is not preserved entirely.

We argue that this currently neglected information is of high value in a long-term preservation strategy of Web 2.0 data and describe our approach for the preservation of wiki contents that is based on Semantic Web technologies. In particular we describe the distributed architecture of our wiki preservation prototype (Urobe) which implements a migration strategy for wiki contents and is based on Semantic Web and Linked Data principles. Further, we present a first vocabulary for the description of wiki core elements derived from a combination of established vocabularies/standards from the Semantic Web and digital preservation domains, namely Dublin Core, SIOC, VoiD and PREMIS.

1. INTRODUCTION

Users of Web 2.0 applications like wikis, blogs or social networking tools generate highly interlinked data of public, corporate and personal interest that are increasingly considered for long-term digital preservation. The term Web 2.0 can be regarded as referring to "a class of Web-based applications that were recognized ex post facto to share certain design patterns", like being user-centered, collaborative and Web-based [6]. Web 2.0 applications are usually based on flexible, lightweight data models that interlink their core elements (e.g., users, wiki articles or blog posts) using hyperlinks and expose these data on the Web for human consumption as HTML. The current practice for the preservation of these data is to treat this layer like a regular Web site and archive the HTML representations (usually by crawling them) rather than the core elements themselves. By this, some irrelevant information (like e.g., automatically generated pages) is archived while some valuable information about the semantics of relationships between these elements is lost or archived in a way that is not easily processable by machines[1]. For example, a wiki article is authored by many different users and the information who authored what and when is reflected in the (simple) data model of the wiki software. This information is required to access and integrate these data with other data sets in the future. However, archiving only the HTML version of a *history* page in Wikipedia makes it hard to extract this information automatically.

Another issue is that the internal structure of particular core elements (e.g., wiki articles) is currently not preserved adequately. Wiki articles are authored using a particular wiki markup language. These simple description languages contain explicit information about the structure of the text (e.g., headings, emphasized phrases, lists and tables, etc.). This internal structure is lost to some extent if digital preservation strategies consider only the HTML version of such articles rendered by a particular wiki software as this rendering step is not entirely reversible in many cases.

In a summary, we state that the current practice for the preservation of Web 2.0 data preserves only one particular (HTML) representation of the considered data instead of preserving the core elements of the respective data models themselves. However, we consider these core elements and their relations crucial for future data migration and integration tasks. In the following we introduce our system Urobe that is capable of preserving the core elements of data that are created using various wiki software.

[1] Cf. http://jiscpowr.jiscinvolve.org/wp/2009/03/25/arch-wiki/

2. UROBE: A WIKI PRESERVATION TOOL

We are currently developing a prototype (Urobe) for the long-term preservation of data created by wiki users. One particular problem when considering wiki preservation is that there exists not one single but a large number of different wiki implementations[1], each using its own wiki markup language. This is what makes a general emulation approach for preserving wiki contents unfeasible as it would require establishing appropriate emulation environments for each wiki software. After further analyzing several popular wiki engines, we have identified the following required components for implementing a long-term, migration-based wiki preservation strategy:

1. An abstract, semantic vocabulary / schema for the description of core elements and their relations stored in a wiki, namely: users, articles and revisions, their contents, links, and embedded media.
2. Software components able to extract these data from wiki implementations.
3. A scalable infrastructure for harvesting and storing these data.
4. Migration services for migrating contents expressed in a wiki markup language into standardized formats.
5. Migration services for the semantic transformation of the meta data stored in this system to newer formats (i.e., services for vocabulary evolution).
6. Software interfaces to existing digital preservation infrastructures using preservation meta data standards.
7. An effective user interface for accessing these data.

2.1. Benefits of Semantic Web Technologies

Urobe is implemented using Semantic Web technologies: It extracts data stored in a wiki and archives it in the form of named RDF graphs [4]. The resources and properties in these graphs are described using a simple OWL[2] vocabulary. Resources are highly interlinked with other resources due to structural relationships (e.g., article revisions are linked with the user that authored them) but also semantic relationships (e.g., user objects stemming from different wikis that are preserved by Urobe are automatically linked when they share the same e-mail address). We decided to make use of Semantic Web technologies for the representation of preserved data and meta data for the following reasons:

Flexibility. The data model for representing the preserved wiki contents is likely to change over time to meet new requirements and it is not predictable at present how this data model will evolve in the future. In this context, modelling the data with the flexible graph-based RDF data model seems a good choice to us: migrating to a newer data model can be seen as an ontology matching problem for which tools and methods are constantly being developed in Semantic Web research [5, 11].

High semantic expressiveness. In order to read and interpret digital content in the future, it is necessary to preserve its semantics. As a consequence of the continuous evolution of data models, knowledge about data semantics disappears quickly if not specified explicitly [11]. To face this problem, we make use of well-defined standardized Semantic Web vocabularies to define the semantics of our data explicitly.

Existing inference support. Inference enables to find relations between items that were not specified explicitly. By this it is possible to generate additional knowledge about the preserved data that might improve future access to and migration of the data.

Expressive query language. One of the key goals of digital preservation systems is to enable users to re-find and access the data stored in such an archive. This often requires a preservation storage to enable complex and sophisticated queries on its data. Data in RDF graphs can be queried using SPARQL, a rich, expressive, and standardized query language that meets these requirements [8].

Furthermore, we decided to publish the archived data as *Linked Data* [2] in order to exchange them between de-centralized components. Linked Data means that (i) resources are identified using HTTP URIs (ii) de-referencing (i.e., accessing) a URI returns a meaningful representation of the respective resource (usually in RDF) and (iii) these representations include links to other related resources. Data published in this way can easily be accessed and integrated into existing Linked Data.

This highly flexible data representation can be accessed via the Urobe Web interface and can easily be converted to existing preservation meta data standards in order to integrate it with an existing preservation infrastructure.

2.2. A Vocabulary for Describing Wiki Contents

We have developed an OWL Light vocabulary for the description of wiki core elements by analyzing popular wiki engines as well as common meta data standards from the digital preservation domain and vocabularies from the Semantic Web domain. The core terms of our vocabulary are depicted in Figure 1. Our vocabulary builds upon three common Semantic Web vocabularies: (i) *DCTERMS* for terms maintained by the Dublin Core

[1]For example, the website http://www.wikimatrix.org/ lists over 100 popular wiki engines.
[2]http://www.w3.org/2004/OWL/

Metadata Initiative[1], (ii) *SIOC* for describing online communities and their data[2] and (iii) *VoiD* for describing datasets in a Web of Data[3].

The vocabulary was designed to be directly mappable to the PREMIS Data Dictionary 2.0 [9]. We have implemented such a mapping and enable external tools to access the data stored in an Urobe archive as PREMIS XML descriptions[4]. This PREMIS/XML interface makes any Urobe instance a PREMIS-enabled storage that can easily be integrated into other PREMIS-compatible preservation infrastructures.

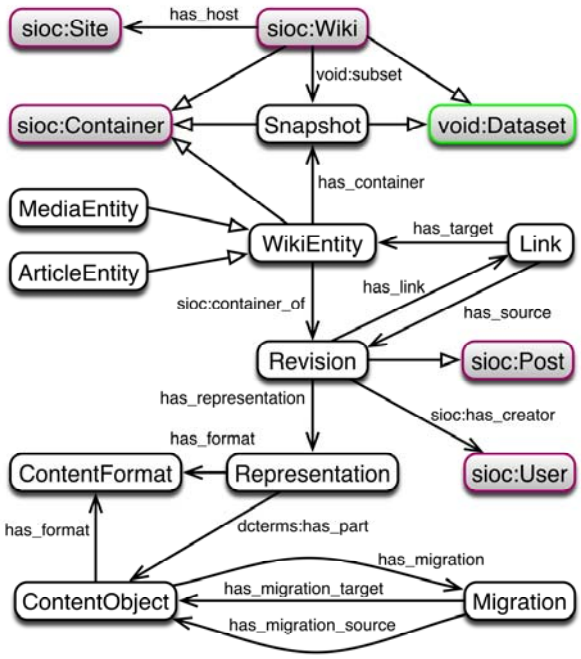


Figure 1. Core terms of the Urobe vocabulary for describing wiki contents. The vocabulary is available at http://urobe-info.mminf.univie.ac.at/vocab.

PREMIS is extensible by design: As RDF graphs can be serialized to XML[5] they are directly embeddable in PREMIS descriptions (using the *objectCharacteristicsExtension* semantic unit). Further, it is possible to describe media objects (i.e., images, videos, documents) that are embedded in wiki pages using appropriate semantic vocabularies like the COMM multimedia ontology [1] (an MPEG-7 based OWL DL ontology that covers most parts of the MPEG-7 standard) or the Music Ontology[6] (that provides a formal framework for describing various music-related information, including editorial, cultural and acoustic information). These descriptions can then be embedded in/mapped to PREMIS descriptions. These meta data could partly be extracted from the object's content itself (e.g., from ID3v2 tags or XMP headers) but also be retrieved directly from the Web of Data (e.g., from the MusicBrainz database, cf. [10]) which could enhance the quality of these meta data considerably.

2.3. Migration of Wiki Articles

Some time ago, the wiki research community started with a first standardization attempt for wiki markup languages[7] which led to a first stable recommendation (Creole 1.0). We therefore decided to implement tools for migrating the source code of wiki articles from their original wiki markup language to Creole 1.0 as soon as they are integrated into our preservation storage. So far we have implemented migration tools for the markup languages of MediaWiki and JspWiki based on components from the WikiModel project[8].

Creole is a wiki markup that contains common elements of many existing wiki engines. However, it is not able to express all specialized elements that are available in the various markup languages[9]. This means that converting wiki articles to Creole is often a lossy migration step. Therefore Urobe additionally preserves the original article source code in its original markup language to enable less lossy migration in the future. However, some loss is unavoidable in such a migration, although it might concern mostly features of minor importance (such as e.g., specialized coloring of table headings or borders around embedded images). If such features have to be preserved, storing the HTML representation of wiki articles is unavoidable. However, even in this case we consider the preservation of a wiki's core elements as beneficial as it enables integration of the data with other data but also direct reasoning on the archived contents.

Further it is notable, that preserving the article source code instead of its rendered HTML version saves a lot of space in a preservation storage: When we compared the raw byte sizes of HTML and plain source code representations of random Wikipedia articles, we found out that the source code representation uses less than 10% of the HTML size in most cases. Thus, the storage requirements for a Wiki archive could be reduced considerably if the mentioned migration loss is considered acceptable in a particular wiki preservation strategy.

As mentioned before, not only the data themselves but also their semantics that are expressed using our OWL vocabulary will have to be migrated in the future. We have not yet developed tools for the migration of our vocabulary, but are confident that this can be

[1] http://purl.org/dc/terms/
[2] http://sioc-project.org/
[3] http://vocab.deri.ie/void/
[4] In compliance to the Linked Data recommendations access to these representations is possible via content negotiation.
[5] http://www.w3.org/TR/REC-rdf-syntax/
[6] http://musicontology.com/
[7] http://wiki.wikicreole.org/
[8] http://code.google.com/p/wikimodel/
[9] An example are mathematical formulas in media wiki, see http://en.wikipedia.org/wiki/Help:Displaying_a_formula.

achieved by using tools and methodologies from ontology matching research.

2.4. Modularized, Distributed Architecture

Urobe is a distributed Web application that comprises three central components:

Proxy components access particular wiki implementations, convert their data to RDF and expose these RDF graphs as Linked Data. Proxies know how to access the data stored by a particular wiki software (e.g., by directly interacting with the database the wiki stores its data in).

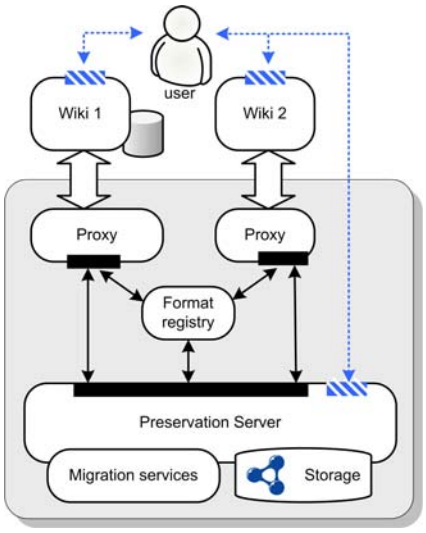


Figure 2. Core architecture of our Urobe prototype. Solid black boxes and arrows denote Linked Data interfaces, dashed blue boxes and arrows denote HTML interfaces.

The format registry stores descriptive and administrative meta data about particular file formats, including descriptions of various wiki markup languages.

The preservation server periodically accesses the proxy components using HTTP requests and harvests all data that were created since the proxy was last accessed. These data are stored in a triple store and interlinked with other data stemming from other wiki instances (e.g., user objects are automatically interlinked when they share a common e-mail address). Such links are then exploited e.g., for data access via the Urobe GUI. A preservation server is able to archive multiple wikis of different types. The internal architecture of the preservation server is influenced by the reference model for an Open Archival Information System (OAIS). Migration services for various object types can be plugged into this server. Currently, object migration is done either immediately after ingestion (for wiki articles) or on demand (for media objects). A preservation workflow component is under development.

The components of Urobe are loosely coupled via HTTP interfaces (cf. Figure 2). This modular architecture and the standardized protocols and formats used by Urobe allow for the easy integration of its components into other applications.

2.5. A Web Interface for Accessing the Urobe Preservation Storage

Human users may access Urobe via an HTML interface (Figure 3) provided by the preservation server component. This interface enables them to search for wiki contents in the Urobe archive using full-text queries and a faceted search approach. Facets for filtering result sets include (i) the wiki(s) the user wants to search, (ii) the time interval the results were created in, (iii) content types and, (iv) the size of multimedia objects. The detail view of articles/media objects presents a timeline of the preserved revisions of this item that indicates all revisions that were created within the search time frame using a different color. Users may navigate to other revisions by simply clicking into the timeline. The original source code of an article as well as all migrated representations are accessible via this screen. A HTML version that is rendered from the preserved Creole source code comprises the default view of an article. Machine actors may further access PREMIS/XML and RDF representations of the stored wiki contents using the Linked Data interface that exposes these data in a machine processable format. The various representation formats are accessible via content negotiation: e.g., when the content type *text/n3* is passed in the Accept header of the HTTP request, Urobe returns a N3-serialized RDF graph describing the respective resource. Urobe also provides a SPARQL endpoint for formulating complex queries over the preservation storage. As future work we further consider to implement a time-based content negotiation mechanism for accessing our preservation storage, as recently presented in [12].

3. CONCLUSIONS

We have formulated requirements and presented a first approach for the digital long-term preservation of wikis, a particular type of a Web 2.0 application. Our approach strongly relies on the adoption of Semantic Web methods and technologies. Wiki contents are modeled using a graph-based data model and their semantics are described using a simple OWL ontology. The advantages of Semantic Web technologies for digital preservation tasks were also recognized by others [7, 8, 3], especially the flexible and extensible way of data representation is considered as beneficial for future data and vocabulary migration as well as for data integration tasks.

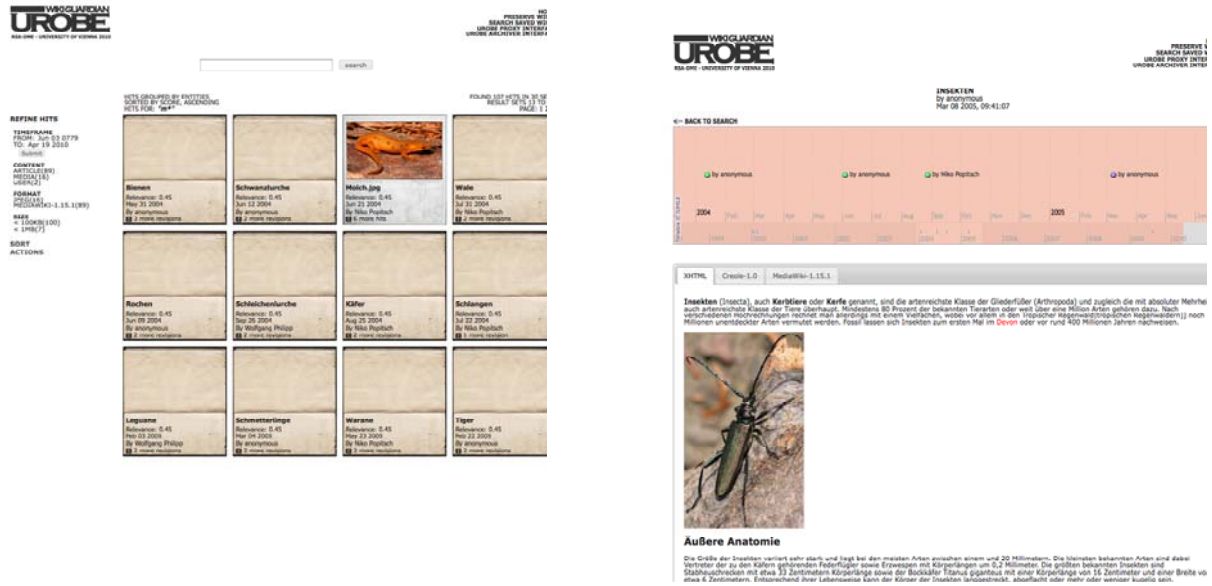
We further contribute a first vocabulary for the abstract description of wiki contents which we consider a precondition for a general wiki preservation strategy. We envision this vocabulary to be continuously improved in the future, which requires algorithms and tools for migrating the data in a Urobe preservation storage to a new vocabulary version. As discussed, we have not yet implemented such a functionality, but due to the strong application of Semantic Web technologies we can benefit directly from the ongoing research in the area of ontology matching.

In the course of the ongoing Urobe project, we aim at extending our vocabulary and implementing support for other semantic vocabularies that are able to capture additional aspects of the preserved data that are of importance in digital preservation, such as context information and provenance meta data.

Finally, our proposed way of exposing the data stored in Urobe as Linked Data enables others to link to these data in a standardized way without compromising their integrity. These externally linked data could then be exploited to harvest additional preservation meta data and ultimately to improve future content migration steps. Further, others could directly benefit from the invariant data in such an archive by being able to create stable links to particular revisions of wiki core elements.

4. AKNOWLEDGEMENT

This research is partly funded by the *Research Studios Austria* program of the Federal Ministry of Economy, Family and Youth of the Republic of Austria (BMWFJ).



(a) Main search screen.

(b) Detail view.

Figure 3. Urobe graphical user interface. The left screenshot shows the main search screen, including the full-text search and the faceted search interface. The right screenshot shows the detail view of a preserved article: the timeline on top of the screen visualizes the revisions of the corresponding wiki article. Below, various representations of the article (XHTML, Creole, original markup) can be accessed.

5. REFERENCES

[1] Richard Arndt, Raphael Troncy, Steffen Staab, Lynda Hardman, and Miroslav Vacura. COMM: Designing a well-founded multimedia ontology for the web. In *Proceedings of the ISWC '07*, pp. 30–43, 2007.

[2] Tim Berners-Lee Christian Bizer, Tom Heath. "Linked data - the story so far," *IJSWIS*, 5(3):1–22, 2009.

[3] Laura E. Campbell. Recollection: "Integrating data through access," In *Proceedings of the ECDL '09*, pp. 396–397, 2009.

[4] Jeremy J. Carroll, Christian Bizer, Pat Hayes, and Patrick Stickler. CNamed graphs, provenance and

trust," In *Proceedings of the WWW '05*, pp. 613–622, 2005.

[5] Giorgos Flouris, Dimitris Manakanatas, Haridimos Kondylakis, Dimitris Plexousakis, and Grigoris Antoniou. "Ontology change: Classification and survey," *Knowl. Eng. Rev.*, 23(2):117–152, 2008.

[6] Mark Greaves. Semantic web 2.0. *IEEE Intelligent Systems*, 22(2):94–96, 2007.

[7] Jane Hunter and Sharmin Choudhury. "Panic: an integrated approach to the preservation of composite digital objects using semantic web services". *Int. J. Digit. Libr.*, 6(2):174–183, 2006.

[8] Gautier Poupeau and Emmanuelle Bermès. "Semantic web technologies for digital preservation : the spar project", In *Proceedings of the Poster and Demonstration Session at ISWC2008*, 2008.

[9] PREMIS Editorial Committee. Premis data dictionary for preservation metadata, version 2.0, 2008. http://www.loc.gov/standards/premis/.

[10] Yves Raimond, Christopher Sutton, and Mark Sandler. "Interlinking music-related data on the web," *IEEE MultiMedia*, 16(2):52–63, 2009.

[11] Christoph Schlieder. Digital Heritage: Semantic Challenges of Long-term Preservation. *submitted to the Semantic Web Journal (SWJ)*, 2010. http://www.semantic-web-journal.net/content/new-submission-digital-heritage -semantic-challenges-long-term -preservation.

[12] Herbert Van de Sompel, Robert Sanderson, Michael Nelson, Lyudmila Balakireva, Harihar Shankar, and Scott Ainsworth. "An HTTP-based Versioning Mechanism for Linked Data," *LDOW2010, Co-located with WWW '10*, 2010.

[13] W3C Semantic Web Activity - RDF Data Access Working Group. Sparql query language for rdf. Technical report, W3C, 2008.

APPROACHES TO ARCHIVING PROFESSIONAL BLOGS HOSTED IN THE CLOUD

Brian Kelly **Marieke Guy**

UKOLN,
University of Bath
Claverton Road, Bath
UK

ABSTRACT

Early adopters of blogs will have made use of externally-hosted blog platforms, such as Wordpress.com and Blogger.com, due, perhaps, to the lack of a blogging infrastructure within the institution or concerns regarding restrictive terms and conditions covering use of such services. There will be cases in which such blogs are now well-established and contain useful information not only for current readership but also as a resource which may be valuable for future generations.

The need to preserve content which is held on such third-party services ("the Cloud') provides a set of new challenges which are likely to be distinct from the management of content hosted within the institution, for which institutional policies should address issues such as ownership and scope of content. Such challenges include technical issues, such as the approaches used to gather the content and the formats to be used and policy issues related to ownership, scope and legal issues.

This paper describes the approaches taken in UKOLN, an applied research department based at the University of Bath, to the preservation of blogs used in the organisation. The paper covers the technical approaches and policy issues associated with the curation of blogs a number of different types of blogs: blogs used by members of staff in the department; blogs used to support project activities and blogs used to support events.

1. BLOG USAGE WITHIN UKOLN

UKOLN is a national centre of expertise in networked information management based at the University of Bath. Our interest in innovation may require staff to use services, such as blogs, which are not provided within our organisation or by our host institution.

Since UKOLN has interests in digital preservation we seek to ensure that we use our experiences to inform best practices on long term access to content held on such services. Such experiences are beneficial in our role in advising UK higher educational institutions on best practices related to use of new and emerging technologies. This paper describes the approaches we have taken and provides advice for other institutions which may have similar concerns.

2. CASE STUDIES

This paper describes three scenarios illustrating differing uses of blogs in UKOLN and highlights the challenges the examples provide regarding the preservation of the contents of blogs.

2.1. The Professional's Blog

The UK Web Focus blog (see http://ukwebfocus.wordpress.com/) was established by Brian Kelly in November 2006. Although there had been some previous experimentation with use of blogs this was the first high-profile blog to be provided by a member of staff and endorsed by JISC (UKOLN's core funding organisation) as a key user engagement and dissemination channel for aspects of UKOLN's work. Since at the time the blog was established neither UKOLN nor the University of Bath provided a blogging platform the WordPress.com service was selected to host the blog.

Since its launch over 750 posts have been published (an average of about four per week) and the blog has attracted over 250,000 user visits (an average of about 240 per day).

This blog supports the author's professional activities and is also written in a personal style which reflects the author's interests and personality. The same is true of Marieke Guy's Rambling of a Remote Worker blog (see http://remoteworker.wordpress.com/).

These two examples illustrate how there may be a degree of uncertainty as to whether the blog posts should be regarded as having institutional or personal ownership.

In light of the popularity and significance of the blog it has been recognised that there is a need to ensure that

best practices are developed in order to minimize the risks associated with use of a third-party service to host the content and the risks of loss of institutional IPR which is managed by the blog author, with no formal mechanism for access by others in the institution and no well-understood levels of accountability for the curation of the content by the author.

In addition to clarity regarding such responsibilities there is a need to identify the tools and processes for curating the blog's content independently from the existing platform and, possibly, ownership.

2.2. The Project Blog

The JISC PoWR (Preservation of Web Resources) project was funded by the JISC and provided by a partnership of UKOLN and ULCC. The project ran from April – November 2008. A WordPress blog was used to support the project work which was hosted by the JISC on their JISC Involve platform (see http://jiscpowr.jiscinvolve.org/).

Content for the blog was provided by staff from the two partner organisations. In order to avoid possible confusions regarding ownership of the content it was agreed that blog posts would be published under a Creative Commons licence and a statement to this effect was provided on the blog.

A decision was made to host the blog on a platform provided by the project's funding body rather than using the host institution of either of the project partners. Although this should avoid the risk of unanticipated changes to terms and conditions for the service we are aware that expected cuts in funding for higher education could result in withdrawal of the service or a failure for the service to be developed. We therefore have an interest in the migration of the content of the blog in the unlikely situation that such changes do occur.

2.3. The Event Blog

UKOLN's Institutional Web Management Workshop (IWMW) is an annual 3-day event. The event provides an opportunity to demonstrate uses of innovative Web technologies. After use of wikis and social networking services in previous years in 2009 the choice was made to use an externally-hosted WordPress.com blog (see http://iwmw2009.wordpress.com/).

In addition to posts from the organisers, speakers and other participants at the event were invited to contribute to the blog. Interviews with participants were also published on the blog both as text and video interviews.

As well as embedded video clips (which are hosted on the Vimeo video sharing service) the blog also provided embedded photographs taken at the event which are hosted on Flickr.

This event blog has given rise to some additional challenges related to the long-term preservation of the content including the ownership of content provided by contributors who do not work for UKOLN, privacy issues related to hosting photographs of participants at the event and the sustainability of the content hosted on other third party services.

3. WHAT ARE THE REQUIREMENTS?

Although three different use cases for organisational blogs have been provided it is not necessarily the case that the same requirements will be needed for the 'archiving' of the blogs. It should be noted that the 'archiving' term is being used to describe ways in which blog content can be migrated to alternative environments in order to satisfy a number of business functions, including the re-creation of the original environment. A number of approaches have been identified which are relevant to our use cases:

Production of a new static master version of the content: This approach is felt to be appropriate for use of project blogs when the project has ceased. The contents of the blog can be migrated as static HTML pages. In order to avoid confusion with multiple copies of the content being available the original blog may have a pointer to the new static resource, possibly with the original content being removed from public view.

Production of a backup version of the content: There may be a need to ensure a backup copy of a blog is available in order to avoid the risks of loss of data if the hosting service is not sustainable or, if as has been seen in the case of the Theoretical Librarian blog (which was hosted at http://theoretical-librarian.blogspot.com/) a blog is removed by the service provider, as illustrated in Figure 1.

Figure 1: Removal of a Blog at Blogger.com

Migration of the rich content to an alternative platform: It may be felt necessary to migrate the contents of a blog to an alternative blogging platform in order to ensure that the blogging characteristics will continue to be available. This might include the migration of a live blog to an alternative platform (which would not normally be described as archiving) but could also involve copying the blog's rich content in order to support data mining or other business processes which may not be possible on the original environment.

Production of a physical manifestation of the content: It may be felt desirable to produce a physical manifestation of a blog, such as a hard copy printout, for various purposes, including marketing purposes or to provide access to the content when online access is not possible.

4. TECHNICAL APPROACHS

4.1. HTML Scraping

The HTTrack offline browsing software (see http://www.httrack.com/) has been used to create copies of the UK Web Focus and IWMW 2009 blogs. This approach is simple to use and requires no special access permissions in order to archive public blogs. However the archived resource is a static Web site and the blog's structure (individual blog posts, comments, etc.) is no longer available as a managed resource.

4.2. Blog Migration

An experiment to migrate the rich content of the blog took place in July 2007 [5]. A blog was created on the VOX platform and the content of the blog was migrated using the host blog's RSS feed. This approach did maintain the structure of the individual posts although comments were lost. However since the migration relied on the host blog's RSS feed this approach is unlikely to be usable for well-established blogs where RSS feeds typically provide access only to recent posts.

An alternative approach is to use the blog service's export functionality and migrate the content to either different blog software or to a platform hosting the same software. This approach has been used to migrate the UK Web Focus blog to another instance of WordPress which demonstrated that not only blog posts and comments could be successfully migrated but also draft posts and embedded objects.

4.3. Processing RSS Feeds

Despite limitations of RSS to provide content for reuse, it is possible on WordPress to provide an RSS feed not just for new posts but also for all views of the blog [3]. This feature is currently being evaluated as a mechanism for migrating blogs if it is not possible to have access to an export file – see [11].

4.4. Production of PDFs

On the second anniversary of the launch of the UK Web Focus blog a PDF version of the blog was created [5]. Although this fails to provide a reusable resource there may be use cases for which this provides an appropriate solution for preserving the content of a blog.

4.5. Physical Manifestation of a Blog

Although the provision of a blog in a physical format (such as a printed book) may appear to be an unusual approach to digital preservation this approach could be of interest for a student or researcher wishing to provide tangible evidence of their blogging output. The Lulu print-on-demand service (see http://www.lulu.com/) is currently being evaluated for the production of hard-copy outputs of our blogs. This will include policy decisions on the content to be published (e.g. should comments be included?).

4.6. Third-Party Web Archiving Services

Commercial Web archiving services such as the UK Web Archive (see http://www.webarchive.org.uk/ukwa/) and Archive-It (see http://www.archive-it.org/) provide an alternative approach to the provision of archives.

The UK Web Archive states that "*If you are the owner of a UK website you are especially encouraged to nominate your own site: this will make the permissions process as straightforward as possible. However, please note that we reserve the right to decide whether to include a site and that for technical reasons we may not be able to archive all sites."* [15]. The JISC PoWR blog was submitted to the UK Web Archive service. Archives of the site were gathered in January, April, July and October 2009 and January 2010 but none of the updates to the blog made between January and July 2010.

Archive-It is a subscription service which has "*95+ partners include: state archives, university libraries, federal institutions, state libraries, non government non profits, museums, historians, and independent researchers*" [4]. Examining the Archive-It service in July 2010 revealed that only one resource from the JISC PoWR blog was available in the archive.

5. POLICY AND RELATED ISSUES

5.1. Blog Policies

In addition to the evaluation of various technical approaches for the migration of blog content we have also implemented appropriate policy statements regarding the ownership of the content, access to the content and rights if the blog author leaves the host institution or if there are changes to the terms and conditions or sustainability of the third party service.

The blog polices for the UK Web Focus and Ramblings of a Remote Workers blog state that:

"*A copy of the contents of the blog will be made available to UKOLN if I leave UKOLN. Note that this may not include the full content if there are complications concerning their party content (e.g. guest blog posts, embedded objects, etc.), technical difficulties in exporting data, etc.)"* and "*Since the blog reflects personal views I reserve the rights to continue providing the blog if I leave UKOLN. If this happens I will remove any UKOLN branding from the blog*" [10].

5.2. Risk of Use of Third Party Services

A risk assessment approach to use of third party services to support UKOLN activities was first used at the IWMW 2006 event when a risk assessment statement was published which provided an assessment of risks and plans for mitigating against such risks [12]. Risk statements have been produced for subsequent events

which ensure that the organisers consider the risks they may be taking and also provides documentation on the third party services which are used.

A framework for assessing the risks of use of third party services has been published which builds on these initial approaches [7].

5.3. Privacy Issues

Possible concerns regarding the publication of photographs of participants at the IWMW 2009 event were identified prior to the event. The event booking form used an approach taken for bookings at recent JISC conferences which stated that photographs would be taken at events. However the event organisers would use their discretion when reusing such photographs. In addition we provided 'quiet area' at the event which was intended for participants who did not wish to be photographed or distracted by the noise of use of laptops [13]. We sought to ensure that photographs used on the blog would not be likely to cause embarrassment. We have also agreed that we will be prepared to remove photographs from services under our control if a rights-holder expresses their concerns if this can readily be achieved.

5.4. Ownership Issues

In order to clarify ownership issues we use Creative Commons licences for our blogs. The UK Web Focus blog contains the following statement:

"*This work is licensed under a Creative Commons Attribution-Noncommercial-Share Alike 2.0 UK: England & Wales License. This licence applies to textual content published by the author and (unless stated otherwise) guest bloggers. Comments posted on this blog will also be deemed to have been published with this licence. Please note though, that images and other resources embedded in the blog may not be covered by such licences.*"

Note that the statement acknowledges the complexities of copyright issues. A risk assessment approach is taken based on ideas described in [8].

6. ARCHIVING APPROACHES FOR THE CASE STUDIES

6.1. The JISC PoWR Blog

Our original intention with the JISC POWR blog was to continue to publish occasional posts related to Web preservation issues but at a significantly lower level. Our aim was to allow the blog to be reused if additional funding became available to continue our work in providing advice on best practices for the preservation of Web resources. However although we were successful in obtaining additional funding this covered a broader area that Web preservation. We therefore felt that it was inappropriate to change the scope of the original blog and have chosen to archive the blog.

The process of freezing the JISC PoWR blog involved carrying out an auditing of use of the blog, with a post published containing a summary of the numbers of posts and comments published, numbers of contributing authors, etc.

An audit of the blog technologies used was also carried out and published. This included details of the WordPress plugins installed and theme and widgets used. We became aware of the value of such audits when, in May 2010 the hosting agency upgraded the platform from WordPress 1 to WordPress 2. A consequence of the update was the loss of the theme, with the look-and-feel reverting to the WordPress default. We need to know which theme we had been using in order to recreate the previous appearance of the blog.

In order to have a better understanding of how the blog was used we created a copy of the blog on the UKOLN Intranet. This will enable us to analyse the contents of the blog using a variety of WordPress plugins which are not available on WordPress.com.

The availability of the backup copy of the blog meant that we could change configuration options which we would not want to do on the live blog. We set the number of RSS items provided to a large number so that the entire contents of the blog posts and comments could be made available via an RSS feed. The RSS feed was used to produce a Wordle word cloud which provides a visualization of the contents of the blog and the comments which have been provided. The RSS feed was also processed by Yahoo Pipes. This enabled the contents of the blog to be processed by an RSS to PDF tool, with a series of PDF files being produced in chronological order (with the capability of applying additional filtering if so desired).

A blog post announcing the "*Cessation of posts to the JISC PoWR blog*" was published in July 2010 which helped to ensure that the status of the blog had been provided to visitors to the blog [2].

The archiving approaches taken to the JISC PoWR is summarised as:

A record of the status of a project blog was taken and published. A rich copy of the contents of the blog was held on a Wordpress blog on the UKOLN Intranet which provides a backup managed within the organisation.

6.2. The UK Web Focus Blog

Periodic copies using a rich XML export of the content of the UK Web Focus blog have been created and used to recreate the blog on a Wordpress installation on UKOLN's Intranet.

The ability to configure the backup blog enables additional management and auditing approaches to be carried out on the blog which cannot be implemented on

the live blog due to the limitations provided on the WordPress.com or to avoid changing the environment for users of the live blog.

The appearance of the blog has been changed so that all posts are displayed on a single (very large) HTML page. The contents of this page has been copied and pasted into an MS Word file and an automatic table of contents has been generated. The blog can then be managed in a similar fashion to conventional MS Word documents.

The numbers of RSS items which can be accessed had been changed on the backup blog to a large number, to enable all posts to be processed using RSS (on the live blog only the most recent 25 items are available by the blog's RSS feed). Yahoo Pipes can be used to process the complete contents of the blog, with the ability to provide a variety of filtering mechanisms. This approach has been used to provided PDF copies of the blog on an annual basis, using the RSS2PDF (http://rss2pdf.com/) service. The selected view of the blog can then be managed in a similar fashion to conventional PDF documents.

In addition to these in-house approaches the blog was also submitted to the UK Web Archive service. However no notification has been received from the service and the blog does not appear to have been retrieved by the service.

The archiving approaches taken to the UK Web Focus is summarised as:

Periodic rich copies of the UK Web Focus blog are taken and installed on the UKOLN Intranet for use in more detailed analyses of the blog. The backup can also be used to avoid loss of the content in cases of a lack of sustainability to the master copy.

6.3. The IWMW 2009 Blog

The IWMW 2009 event blog was used in the run-up to the event, during the event and shortly after the event had finished when a number of posts were published after the event summarising the feedback received.

In order to provide clear termination of the blog a post was published which announced its closure [1] in line with advice on best practices for closing blogs published [14].

However since IWMW is an annual event we recognised that we may wish to publish occasional posts linking to the forthcoming event. Since the blog can provide marketing benefits, with links likely to help enhance Google ranking it has been decided that the blog will continue to be hosted on WordPress.com, though with some minor changes:

- A sidebar widget ensures that the status of the blog is clear.
- A widget provides links to key resources related to the event.
- Widgets providing access to dynamic content, such as live Twitter feeds, have been removed.

In the preparation work for the archiving the blog we observed that a number of posts contained embedded objects (such as video clicks hosted on the Vimeo.com service) which did not include a link to the object on the remote service. Since we realized that loss of the embedding mechanism (which is a configurable option in WordPress) would result in loss of the embedded object and no information being provided on the location of the hosted video clips we edited the posts to included a link to the object on the external service as illustrated in Figure 2 (taken from http://iwmw2009.wordpress.com/2009/08/07/take-aways/):

> Note that these three videos are hosted on Vimeo and can be accessed directly at:
> * http://vimeo.com/5976384
> * http://vimeo.com/5976404
> * http://vimeo.com/5976472

Figure 2: Links to embedded objects

We have decided not to keep an XML archive of the blog content since we feel the risk of loss of the content is small and there will be no serious consequences if the content is lost. However we have used WinHTTrack to keep a static copy of the blog which is stored on the UKOLN Intranet.

Figure 3: Closure of the IWMW 2009 Blog

We have also published a static page on the blog which summarises these policies (see http://iwmw2009.wordpress.com/status-of-this-blog/).

An illustration of the home page is shown in Figure 3 with the key features highlighted.

The archiving approaches taken to the IWMW 2009 blog is summarised as:

A static copy of the IWMW 2009 blog is available on the UKOLN Intranet. The backup can also be used in case of a lack of sustainability to the master copy.

7. IDENTIFICATION OF GOOD PRACTICE

The work in understanding appropriate solutions for our archiving of professional blogs hosted in the Cloud has helped us to identify appropriate practices which may be particularly relevant for funding bodies who wish to ensure that project-funded activities which make use of blogs provided by third parties implement appropriate approaches for ensuring that the content provided on such blogs does not disappear unexpectedly.

The checklist we have developed includes the following steps:

Planning: Preparation for archiving blogs should begin before the blog is launched. A blog policy can held to clarify the purpose of the blog and its intended audience.

Clarification of rights: A copyright statement covering blog posts and comments can also minimise the legal risks in archiving the blog.

Monitoring of technologies used: Information on the technologies used to provide the blog, including blog plugins, configuration options, themes, etc. can be useful if a blog environment has to be recreated.

Auditing: Providing an audit of the size of the blog, numbers of comments, usage of the blog, etc. may be useful in helping to identify the value of a blog and in ensuring that interested parties are aware of how well-used the blog was.

Understanding of costs and benefits: The audit should help to inform the decision-making processes regarding the effort which needs to be taken for the selected blog archiving strategy.

Identification and implementation of archiving strategy: The appropriate blog archiving strategy needs to be selected. As illustrated in the case studies this could include 'freezing' a blog on the external service, with an organisational backup copy (in a variety of formats) or the continuation of an active blog, with a backup copy of taken in case of unexpected data loss.

Dissemination: It will be desirable to ensure that end users are aware of the existence of an archived copy. Ideally such information will be made publicly available. The summaries of the approaches taken in the three case studies illustrate that such dissemination work need not be time-consuming to implement.

Learning: During the planning, auditing, selection and implementation of appropriate archiving strategies there are likely to be lessons learnt (such as, in the case of the IWMW 2009 case study the need to include links to external services and not just embed the objects). Such experiences should be used to inform subsequent blogging practices.

Organisational Audit: There is a likely to be a need to carry out an organisation audit of use of blogs held on third party services which may be at risk. Such an audit should initially identify (a) location of such blogs; (b) their purpose(s); (c) the owner(s) and (d) their perceived importance. This information should help to inform decisions on the archiving strategies, along the lines described in this paper.

8. CONCLUSIONS

This paper has reviewed the approaches which have been taken to facilitating long-term access to blogs hosted in the Cloud which are used to support professional activities.

The need to ensure that preservation policies are developed and implemented by JISC-funded projects has been described in [9]. Since many of the blogs provided by JISC-funded development projects may be hosted on third-party services there is a need to document and share the variety of possible technical approaches to the migration of content and related policy issues.

The approaches which have been described seek to address the difficulties which organisations are likely to experience in adopting similar approaches, including the potential difficulties of motivating content providers of the need to address such preservation issues and the limited resources which is likely to be available to implement such practices.

9. REFERENCES

[1] Guy, M. *Last Orders at the IWMW2009 blog*, IWMW 2009 blog, 12 August 2009, <http://iwmw2009.wordpress.com/2009/08/12/last-orders-at-the-iwmw2009-blog/> 2009

[2] Guy, M. *Cessation of posts to the JISC PoWR blog*, JISC PoWR blog, 19 July 2010, <http://jiscpowr.jiscinvolve.org/wp/2010/07/19/cessation-of-posts-to-the-jisc-powr-blog/> 2010

[3] Hirst, A. *Single Item RSS Feeds on WordPress blogs: RSS For the Content of This Page*, OUseful blog, 8 July 2009, <http://blog.ouseful.info/2009/07/08/single-item-rss-feeds-on-wordpress-blogs-rss-for-the-content-of-this-page/> 2010

[4] Internet Archive, *Archive-It*, <http://www.archive-it.org/public/faq.html> 2010

[5] Kelly, B. *A Backup Copy Of This Blog*, UK Web Focus blog, 19 April 2007,

<http://ukwebfocus.wordpress.com/2007/07/19/a-backup-copy-of-this-blog/> 2009

[6] Kelly, B. *The Second Anniversary of the UK Web Focus Blog*, UK Web Focus blog, 31 October 2008, <http://ukwebfocus.wordpress.com/2008/10/31/the-second-anniversary-of-the-uk-web-focus-blog/> 2008

[7] Kelly, B., Bevan, P., Akerman, R., Alcock, J. and Fraser, J. *Library 2.0: Balancing the Risks and Benefits to Maximise the Dividends*, Program (2009), Vol. 43, No. 3, pp. 331-327. <http://opus.bath.ac.uk/15260/> 2009

[8] Kelly, B. and Oppenheim, C. *Empowering Users and Institutions: A Risks and Opportunities Framework for Exploiting the Social Web*, CULTURAL HERITAGE online conference, Florence, 15-16th December 2009. <http://opus.bath.ac.uk/17484/> 2009

[9] Kelly, B. *The Project Blog When The Project Is Over*, Web Focus blog, 15 March 2010, <http://ukwebfocus.wordpress.com/2010/03/15/the-project-blog-when-the-project-is-over/> 2010

[10] Kelly, B. *Blog Policies*, UK Web Focus blog, <http://ukwebfocus.wordpress.com/blog-policies/> 2010

[11] Slideshare. *UK Web Focus Blog Posts 2009*, <http://www.slideshare.net/lisbk/uk-web-focus-blog-posts-2009> 2009

[12] UKOLN. *Risk Assessment For The IWMW 2006 Web Site*, <http://www.ukoln.ac.uk/web-focus/events/workshops/webmaster-2006/risk-assessment/> 2006

[13] UKOLN. *Quiet Area, IWMW 2009*, <http://iwmw.ukoln.ac.uk/iwmw2009/quiet/> 2009

[14] UKOLN, *Closing Down Blogs*, Cultural Heritage briefing document no. 81, March 2010, <http://www.ukoln.ac.uk/cultural-heritage/documents/briefing-81/>, 2010

[15] UK Web Archive, *FAQ*, <http://www.archive-it.org/public/faq.html#605> 2010

DIGITAL PRESERVATION: THE TWITTER ARCHIVES AND NDIIPP

Laura E. Campbell **Beth Dulabahn**

The Library of Congress
Office of Strategic Initiatives
101 Independence Ave, SE
Washington DC 20540

ABSTRACT

On April 14, 2010, the Library of Congress and Twitter made the joint announcement that the Library would receive a gift of the archive of all public tweets shared through the service since its inception in 2006. The media and community response was tremendous, raising many questions about how the Library would be stewarding and providing access to the collection. There are many issues to consider, from the technical mechanisms of transfer to the Library and the ongoing updates to the archive, to curatorial policies, to planning for a new type of research access to a Library collection. The Twitter archive joins a number of born-digital collections at the Library. This year the National Digital Information Infrastructure and Preservation Program (NDIIPP) celebrates a decade of digital preservation actions and discovery, working with a network of over 170 partners resulting in over 200 terabytes of all types of important digital content. Collectively, we have built lasting relationships, helped facilitate natural networks within the overall NDIIPP network, and tested new tools and services that support the work of the partners in the network. We will rely on the collective wisdom of our partners as we grapple with the challenges of curating and serving a digital collection as rich and diverse as the Twitter archive.

1. THE TWITTER ARCHIVES ACQUISITION

Twitter is a microblogging service that enables users to send and receive messages of up to 140 characters, called "tweets." Users share their tweets and others follow tweets in a social network environment. Worldwide, Twitter processes more than 50 million tweets per day and this number is growing exponentially.

On April 14, 2010, the Library of Congress and Twitter made the joint announcement [8,10] that the Library would receive a gift of the archive of all public tweets shared through the service since its inception in 2006. The media and community response to the announcement -- simultaneously tweeted and blogged -- was tremendous, beginning a very public conversation about what the Library would receive and how the Library would be stewarding and providing access to the collection.

Although to many it seems an incongruous acquisition for the Library, the Library holds a wide range of materials in many formats, and collects groups of items as well as individual items. With the receipt of the Twitter archive, the Library continues its long tradition of collecting and preserving personal stories, such as the "man on the street" interviews after Pearl Harbor; personal letters and diaries collected for the Veterans History Project; and conversations between family members preserved in StoryCorps1. Twitter forms part of the historical record of communication in the twenty-first century, capturing news reports, events, and social trends. Minute-by-minute headlines from major news sources such as Reuters, The Wall Street Journal and The New York Times are pushed to Twitter. At the same time, it serves as a platform for citizen journalism with many significant events being first reported by eyewitnesses. It is frequently cited as an important unfiltered record of important events such as the 2008 U.S. presidential election or the "Green Revolution" in Iran.

The Library also has a long history of enriching its collections through donations. The Twitter archive is a gift from Twitter; the agreement is openly available online.[2] After the Twitter announcement, Greg Pass, Twitter's vice president of engineering, said: "We are pleased and proud to make this collection available for the benefit of the American people. I am very grateful that Dr. Billington and the Library recognize the value of this information. It is something new, but it tells an amazing story that needs to be remembered." [8]

[1] Story Corps is available at http://storycorps.org.

[2] The agreement between the Library of Congress and Twitter is available at:
http://blogs.loc.gov/loc/files/2010/04/LOC-Twitter.pdf

For its potential value, the size is relatively small - approximately 5 terabytes for all public tweets from 2006 to early 2010. This makes it considerably smaller than the Library's other web archives[3], which comprise more than 170 terabytes of web sites, including legal blogs, topical and event archives, election campaigns for national office, and websites of Members of Congress.

2. THE ARCHIVE AS CASE STUDY

In addition to providing access to the archive, the Library also sees this acquisition as an opportunity for expanding external collaborations with digital preservation partners. The Library has worked with many of its partners to develop and test mechanisms for content transfer, and does not anticipate any significant problems in the actual transfer of the archive from Twitter to the Library. In terms of curation, however, the Twitter archive pushes the limits of traditional models of curation. The contents of the archive cross virtually all of the subject areas within the Library, making it difficult to assign any division sole custodial responsibility. The Library's research and education partners in this effort - scholars, historians, librarians, archivists and scientists - provide unique perspectives when thinking creatively about digital curatorial responsibility, user access, and Library support and services.

Dr. James Billington, the Librarian of Congress, agrees that the benefit is not only to the American people, but to the Library's goal to gain experience in best practices and procedures in support of its collections. Dr. Billington said: "The Library looks at this as an opportunity to add new kinds of information without subtracting from our responsibility to manage our overall collection. Working with the Twitter archive will also help the Library extend its capability to provide stewardship for very large sets of born-digital materials." [8] The Twitter archive will serve as a case study for the management and preservation of a large corpus of digital data.

3. PRIVACY ISSUES

Twitter is donating an archive of what it determines to be public. Alexander Macgillivray, Twitter's general counsel, as quoted in the New York Times [9], said, "From the beginning, Twitter has been a public and open service." Twitter's privacy policy states: "Our services are primarily designed to help you share information with the world. Most of the information you provide to us is information you are asking us to make public." Under the Twitter terms of service, users give Twitter the right to archive tweets.

There will be at least a six-month window between the original date of a tweet and its date of availability for research use. Private account information and deleted tweets will not be part of the archive. Linked information such as pictures and websites is not part of the archive, and the Library has no plans to collect the linked sites. Moreover, the Library does not expect to make the collection available online in its entirety. Use will be limited to non-commercial private study, scholarship, or research.

The Library understands there are concerns about privacy issues and is sensitive to those concerns. The Library has a long history of respecting sensitive information related to its research collections and will be mindful of those concerns as it develops its access plans. Periodic public communications about the archive will set expectations for privacy and access.

4. RESEARCH USE AND ACCESS

The collections will be made available to non-commercial researchers and to Library staff for internal use. Details about researcher access policies are being developed with input from curators across the Library, taking into account principles that have been applied to existing collections. While wanting to stay consistent with the philosophy that has governed the use of its physical collections, the Library recognizes that a digital collection such as the Twitter archive presents a number of new challenges, and is exploring how to accommodate not only a wide range of research uses, but also a geographically diverse set of users.

Viewed in the aggregate, the Twitter collection can be a resource for current and future researchers to study and understand contemporary culture and political, economic and social trends and topics. The Library has assembled a set of use cases from scholarly publications, news publications and blogs, social scientists, and librarians. The use cases drive the creation of archival policy requirements related to search, access, privacy, and preservation.

For historians, Twitter provides direct witness accounts of events in real time. It also serves as a virtual timeline of communications about events, people and places. This provides an enormous amount of raw unmediated primary source material for historical research.

Daniel J. Cohen, an associate professor of history at George Mason University and co-author of a 2006 book, "Digital History", as quoted in the New York Times [9], said that "Twitter is tens of millions of active users. There is no archive with tens of millions of diaries".

The Twitter archive could be used to study empirically how individuals reacted to a particular historical event. If Twitter existed on on September 11, 2001, the American people would have a real-time chronicle of what people were thinking and feeling on that day. Today, it is a popular environment for "first-

[3] The Library of Congress Web Archives are available at: http://lcweb2.loc.gov/diglib/lcwa/html/lcwa-home.html.

on-the-scene" news reports. In February 2008, the earthquake in the United Kingdom was reported on Twitter at least 35 minutes before it was reported in the mainstream press. Later that year, Twitter was used by eyewitnesses of the Mumbai terror attacks:

> *"Hospital update. Shots still being fired. Also, Metro cinema next door," twittered Mumbaiattack;*
>
> *"Mumbai terrorists are asking hotel reception for room #s of American citizens and holding them hostage on one floor", twittered Dupreee.*

Twitter can be compared to earlier sources of personal information such as diaries and letters. Many of the earlier sources contain mundane or trivial pieces of information that, in aggregate, can tell a detailed and authentic story about everyday life that is difficult to find elsewhere. David Ferriero, the Archivist of the United States, points out that historians often find value, sometimes unanticipated, in what others may see as mundane details of our lives and what they might say about our culture [2]. Paul Freedman, a professor of history at Yale University, agrees. Freedman was quoted in Slate as saying: "Historians are interested in ordinary life." [1] Only time will tell its value. It could be that Twitter content may be studied by future scholars in the same way that the graffiti of Pompeii is being studied by current scholars.

Social scientists are similarly interested in using the collection to study trends and patterns, such as social networks which are of interest to a wide range of disciplines from anthropology to political science, management science, sociology and communications. A 2010 study of Twitter use by social science researchers revealed that Twitter was ranked in the top three services used by researchers to spread information. [6] Researchers are increasingly interested in studying scientific networks and the spread of information inside and outside the scientific community.

Researchers may also study communities that drum up support using Twitter. In June of 2009, political dissidents in Iran used Twitter to voice opposition while Iranian newspapers were heavily censored.[4] Future researchers may choose to extract these archived tweets, using time and location data, in order to draw conclusions about the opinions and attitudes of the period.

Researchers and research organizations are excited at the prospect of exploring the Twitter's rich and varied data in the aggregate, especially with newer data mining and social graph analysis tools that can reveal trends[7]. It is always a challenge to predict how the archive might be interrogated, so user feedback and requests will be collected in the coming months and years to help the Library investigate how its can potentially expose its collections as data. The Library may also enter into technical partnerships with external agencies and organizations to develop search and visualization tools for use with the archive. The Library is currently involved in such a partnership with Stanford University, called the Computational Approaches to Digital Stewardship Project[4], which is focused on new tools for the discovery of digital collections.

5. MANAGING THE ARCHIVE

While the primary focus of the Twitter gift is the retrospective archive - tweets from 2006 to 2010 - the Library and Twitter are working on a mutually-agreeable form for incremental updates. The terms of the gift agreement specify that the Library not make available any tweet less than six months old. Therefore, the technical committee responsible for the ongoing archive must develop a framework for receipt, ingest and management that considers this six month hold.

There are interesting issues to be addressed in the areas of receipt, ingest and management going forward:

- The number of tweets per day has been increasing significantly. This means that each incremental update will be significantly larger as we move forward.
- The terms of the gift require that the Library not make available any tweet that is less than six months old. This means that the incremental update receipt and ingest process needs to take this into account.
- The terms of the gift include only public tweets. This means that the incremental update receipt and ingest process needs to take this into account.
- What practices will be followed to verify that the data received by the Library is the same as the data sent by Twitter?
- What practices will be followed to ingest and store the large number of tweets? Will those practices be affected by the method of update, or by changes in the fields or format of future tweets?
- What processing will the Library be performing for management of the tweets or to make them available to researchers?
- What kinds of services will the Library offer to researchers and how will those affect the management of the tweets?

The Library expects to identify and analyze options that would address these issues. The Library may explore a multi-stage process for receipt that optimizes the processing flows at each stage. A multi-stage approach would also allow for each stage to be configured and tuned for best resource use and flexibility for changes in the volume and/or data. For example, the Library could establish a local (or remote) isolated staging area for receipt of update files or streams. One or more processes could then perform any required verification, processing

[4] http://cads.stanford.edu/.

and adding public tweets older than six months to the data set available for researcher use. The Library expects to explore and test technical options that could make up feasible implementation solutions.

The Library is looking forward to expanding its capabilities to take in, make available, and preserve creative content for current and future generations. The Twitter gift provides an opportunity to make progress that we hope will also benefit other institutions and partners that are addressing some of these same issues.

6. NATIONAL DIGITAL INFORMATION INFRASTRUCTURE AND RESERVATION PROGRAM

The Twitter archive is just one of many born-digital collections that the Library has brought under its stewardship, part of a long history of working with digital content. In 2010 the National Digital Information Infrastructure and Preservation Program (NDIIPP) is celebrating ten years of digital preservation actions and discovery working with a network of over 170 partners resulting in over 200 terabytes of all types of important digital content. Collectively, we have built lasting relationships, helped facilitate natural networks within the overall NDIIPP network, and tested new tools and services that support the work of the partners in the network. The collaboration today has federal and state government partners, commercial content partners, service providers, library and archival institutional partners and international partners. We proudly consider the work of the last decade a true collaboration that reflects enormous transformation in the way libraries will work in the future.

A new outgrowth of NDIIPP is the National Digital Stewardship Alliance (NDSA), a collaborative effort among government agencies, educational institutions, non-profit organizations, and business entities to preserve a distributed national digital collection for the benefit of citizens now and in the future. The NDSA is an inclusive organization that will focus on shared work toward common community goals.

We plan to draw on our partners for assistance in technical approaches to supporting a digital archive of this size, richness, and complexity.

7. REFERENCES

[1] Beam, Christopher. “How future historians will use the Twitter archives.” *Slate*. Web, April 20, 2010. July 11, 2010. <http://www.slate.com/id/2251429>

[2] Ferriero, David. “Tweets: What we might learn from mundane details*.” AOTUS: Collector in Chief*. Web. April 16, 2010. July 11, 2010. <http://blogs.archives.gov/aotus/?p=172>.

[3] Margot Gerritsen– private correspondence to Laura Campbell.

[4] Grossman, Lev. “Iran Protests: Twitter, the Medium of the Movement.” *Time Magazine*. June 17, 2009. Web. uly 11, 2010. <http://www.time.com/time/world/article/0,8599,1905125,00.html>

[5] Haddadi, Meeyoung Cha Juan Antonio Navarro P´erez Hamed. *Flash Floods and Ripples: The Spread of Media Content through the Blogosphere*. Association for the Advancement of Artificial Intelligence, 2009.

[6] Letierce, Julie; Passant, Alexandre; Decker, Stefan Breslin, John G. *Understanding how Twitter is used to spread scientific messages.* Web Science Conference 2010, April 26-27, 2010 Raleigh NC.

[7] McLemee, Scott. “The Mood is the Message.” *Inside Higher Ed*. June 30, 2010. Web. July 11, 2010. <http://www.insidehighered.com/views/mclemee/mclemee296>

[8] Raymond, Matt. "Twitter Donates Entire Tweet Archive to Library of Congress.” Library of Congress, April 15, 2010. Web. July 11,2010. <http://www.loc.gov/today/pr/2010/10-081.html>

[9] Stross, Randall. “When History is Compiled 140 Characters at a Time.” *The New York Times:* April 30, 2010. <http://www.nytimes.com/2010/05/02/business/02digi.htm>

[10] Twitter. “Tweet Preservation.” Twitter, April 15, 2010. Web. July 11,2010. <http://blog.twitter.com/2010/04/tweet-preservation.html>

[11] Wasserman, Stanley; Galaskiewicz ,Joseph. Advances in social network analysis: research in the social and behavioral sciences. Sage Publishing, 1994.

TWITTER ARCHIVING USING TWAPPER KEEPER: TECHNICAL AND POLICY CHALLENGES

Brian Kelly
UKOLN,
University of Bath,
Bath, UK

Martin Hawksey
JISC RSC Scotland North & East, Edinburgh's Telford College,
Edinburgh, UK

John O'Brien
3930 Rolling Hills Drive,
Cumming, GA 30041,
USA

Marieke Guy
UKOLN,
University of Bath,
Bath, UK

Matthew Rowe
Department of Computer Science, University of Sheffield,
Sheffield, UK

ABSTRACT

Twitter is widely used in a range of different contexts, ranging from informal social communications and marketing purposes through to supporting various professional activities in teaching and learning and research. The growth in Twitter use has led to recognition of the need to ensure that Twitter posts ('tweets') can be accessed and reused by a variety of third party applications.

This paper describes development work to the Twapper Keeper Twitter archiving service to support use of Twitter in education and research. The reasons for funding developments to an existing commercial service are described and the approaches for addressing the sustainability of such developments are provided. The paper reviews the challenges this work has addressed including the technical challenges in processing large volumes of traffic and the policy issues related, in particular, to ownership and copyright.

The paper concludes by describing the experiences gained in using the service to archive tweets posted during the WWW 2010 conference and summarising plans for further use of the service.

1. ABOUT TWITTER

Twitter has been described as a 'micro-blogging' service. It provides blogging functionality, but the blog posts (often referred to as 'tweets') are restricted to 140 characters. Although this constraint may appear to provide a severe limitation on use of Twitter in an educational and research content, in practice the ease of creating tweets (without the need for the individual to spend time and mental energy in composing their thoughts and perhaps having ideas reviewed by others or checked by an editorial board) has given rise to Twitter being used to support educational and research activities in ways which had not previously being considered. Twitter's popularity has been enhanced by the ability to publish material on a wide range of devices and in particular mobile devices where the 140 character constraint is less of an issue for the small (or virtual) keyboards to be found on such devices.

2. HOW TWITTER IS BEING USED

The growing importance of preservation of Twitter content is illustrated by two examples of existing use to support education and research.

2.1. Supporting Events

Twitter has been used to support a number of high profile events in the UK's higher education community. It has been used by delegates, both physical and virtual, to engage in discussion and disseminate resources. The international ALT-C 2009 conference, which was held over 3 days in September 2009 generated 4.317 Twitter posts from 633 contributors using the #altc2009 hashtag [9]. The JISC's recent annual one-day conference was held in April 2010 generated 2,801 tweets from 479 contributors using the #jisc10 hashtag [10].

UKOLN's annual Institutional Web Management Workshop (IWMW) has made use of networked technologies to support events since an IRC channel was used to support discussions at IWMW 2005. In recent years Twitter has been used; at the IWMW 2009 event

the #iwmw2009 hashtag was used with 1,530 tweets posted from 170 contributors [11]. In addition to the identification of a recommended hashtag for the event the organisers set up a dedicated Twitter account to send announcements as well as providing an official commentary of some of the sessions. One of the plenary speakers at the IWMW 2009 also used Twitter in an innovative way, abandoning use of PowerPoint or other presentation tools, instead simply speaking and responding to tweets from the audience (and a remote audience who were following the event's hashtag) which were displayed on a large screen in the auditorium [14].

2.2. Captioning Video

Work on the use of Twitter as a method for captioning videos has been ongoing since 2009 [7]. The core concept is to convert tweets posted during a live event into a compatible caption file format which then can be replayed with audio or video clips. The development of Twitter based captions has mirrored the increasing use of Twitter to support events and provides a means for delegates to replay archived audio and video recordings with the real-time stream of tweets, in essence allowing users to replay conference sessions augmented with the original backchannel communication. More recently this work has been extended allowing users to generate and play subtitles for on demand television services such as the BBC's iPlayer [2] and political speeches [3].

The software has been developed to use an event hashtag not only to generate subtitles but also to use this resource to allow users to search within the associated media asset. This development opens up the use of Twitter subtitles as a tool to support the increasing popular use of lecture capture in education; that is as well as students being able to replay a captured lecture they can also view the back channel discussions [4].

2.3. Observing Political Debate

Twitter provides real time information about a diverse range of topics, in essence allowing users to be harnessed as social sensors. For instance work by Sakaki [20] has found that Twitter users in Japan could be used as sensors to detect earthquakes by observing their tweets and the location where they were published.

Politics is one of the most discussed topics on Twitter, allowing public sentiment to be gleaned from the analysis of tweets. For instance, work by [1] performed sentiment analysis over a corpus of tweets archived in the run up to the US presidential election of 2008. This allowed public reaction to policy decisions and speeches to be gauged without the need for exhaustive polling.

The archival of tweets discussing politics provides a useful backdated corpus which can be used to explore public reaction and sentiment. Observations made over such data could in turn allow informed future policy decisions to be made.

2.4. Additional Uses

We have shown examples which demonstrate how Twitter is being used within the Higher Education sector. The use of Twitter at events illustrates reasons why tweets should not be regarded as possible value only at the time they were posted as increasingly we might expect to see tweets being analysed after an event in order to inform the evaluation of the event.

Additional reasons why there is a need to ensure that tweets should be made available for reuse include:

- To allow for analysis of Twitter communities e.g. analysis of Twitter spammers [5].
- Analysis of tweets associated with a hashtag used to support sharing and community-building across development programmes such as the JISC Rapid Innovation programe [19] as described at [6].
- Reputation management for both organisations and individuals.
- Personal interests: e.g. to enable a Twitter user to be have an answer to the question "What was I saying when I was young?"

3. WHY THE NEED FOR AN ARCHIVING SERVICE?

Since Twitter provides a search interface to its service it may not be apparent why a third party service is needed to provide an additional archive of tweets.

A key reason which has led to the development of a number of Twitter archiving services is due to limitations of the Twitter search API which provides access only to recently posted tweets. Current documentation from the Twitter Search API states that searches are limited to 1,500 individual tweets from the last 7 days [15]. Consequently as well as not being able to access tweets older than 7 days, the complete timeline for popular 'trending' topics are also not available.

To date Twitter has been designed to facilitate development of third party services around its service, avoiding the need for new features having to be provided by Twitter. Useful additional features which have evolved include statistical analyses, enhanced search capabilities, the ability for end users to manage their collection of Twitter archives and export the data in a variety of formats. As well as such services aimed at end users Twitter archiving services can themselves provide APIs which allow them to be used to provide additional services to developers.

An awareness of the importance of Twitter archiving to the UK's tertiary education community led to the JISC exploring options for a Twitter archiving service.

4. TWITTER ARCHIVING SERVICE OPTIONS

A variety of archiving services for tweets are available. These include WTHashtag (archives tweets for specified hashtags – see http://wthashtag.com/); BackUpMyTweets (used by a Twitter user to provide a

backup of their own tweets – see http://backupmytweets.com/) and Twapper Keeper (see http://twapperkeeper.com/).

These services are based on the Twitter APIs. It would be therefore possible to develop a new service for archiving tweets. However, as identified in the JISC 2010 Strategy "*The New JISC Strategy comes amid serious economic recession in the UK*" [8]. Despite such economic concerns, as the JISC Strategy document identifies: "*Cloud computing offers flexibility and, where the business case is done carefully and accurately, can offer considerable savings by avoiding the cost of owning and large computer facilities and the associated running costs*".

Such strategic considerations provided the context in which, instead of commissioning development of a prototype which, if successful, would be expected to evolve into a national service, it was felt that a more cost-effective development route would be to fund developments of an existing service to ensure that needs of the UK's higher educational sector were addressed.

Following negotiations the JISC agreed to fund developments to Twapper Keeper for a 6 month period from April 2010, with UKOLN, a JISC-funded centre of expertise in digital information management, providing the project management for this work.

The Twapper Keeper blog was used to announce the development work and invite suggestions on developments [17]. The suggestions which were received included:

- Ability to group collections of archives.
- Ability to delete tweets from the Twapper Keeper's archives.
- Ability to opt-out of being archived.
- Provision of APIs to the Twapper Keeper service.
- Access to archives provided in multiple formats (e.g. RSS, Atom and JSON).

There is a need to ensure that the JISC investment in this development work provides a sustainable service. This challenge is nothing new – project-funded work carried out in UK higher educational institutions cannot be guaranteed to result in the delivery of sustainable services. However funding of an external service based outside the UK, while not new, does necessitate that careful attention is taken to not only the development work itself but also the sustainability of the service after the development work is over.

The approaches which are being taken in the development work include:

- **An open approach to developme**nt: to gain buy-in from the user community and ensure developments reflect the communities' needs.
- **Migration of the platform**: to a more stable platform to ensure that the service can cope with the anticipated growth in use and traffic.
- **Open sourcing components**: which would allow the service to be replicated if this was felt to be necessary.
- **Open content for documentation**: through use of Creative Commons licences for the project blog, technical documentation, FAQs, etc.

5. CHALLENGES

Following the gathering of the user requirements for developing the Twapper Keeper service we have prioritised the requirements and developed an implementation plan.

The following technical and policy challenges in implementing the user requests have been identified:

5.1. Technical Issues

Due to the rapid adoption of the Twitter service, the Twitter API and ecosystem continues to evolve. This requires services such as Twapper Keeper to continue to develop and align with the Twitter technical and policy changes. For example, recent changes to the way Twitter recommends tweet data be accessed and consumed from RESTful search service to Streaming APIs has shifted the load of processing tweets to Twapper Keeper forcing the system to take on more burden when trying to store large numbers of tweets.

Another example is alignment with policy requirements for deletion of tweets. Twitter requires third party systems leveraging the API to delete tweets when a user deletes a tweet from Twitter, which requires a delete event to be sent from Twitter to the third party system. However limitations in the current RESTful search / timeline APIs and the Streaming API for tracking keywords results in no deletion events being sent to the Twapper Keeper service. Therefore, to overcome this from a policy perspective the Twapper Keeper service requirements users to inform the service if they want their tweets to be deleted.

A final example of the technical challenges is that changes continue to take place in the Twitter ecosystem, as highlighted by the upcoming plan to drop support for Basic HTTP Authentication with Twitter API REST services and the requirement for services to migrate to use OAuth. The ability of the Twapper Keeper service to manage such changes in accessing data held by the Twitter service enables third party services to avoid the need to make modifications to their service when the Twitter backend access mechanisms are changed.

5.2. Policy Issues

The ability for a user to delete their tweets from the Twapper Keeper archive has been requested as well as users being able to opt-out completely from the service. This request reveals uncertainties regarding the copyright status of Twitter posts and the ways in which third party services should address issues of ownership and related management issues.

The Twitter terms of service state that "*You retain your rights to any Content you submit, post or display on or through the Services*" [16]. This could be

interpreted to mean that it would not be possible for tweets to be harvested by others without permission of the owner. However this is clearly not a scalable solution as can be seen by the popularity of Twitter archiving services (including the announcement that the US Library of Congress is to archive tweets [13]). However rather than disregarding concerns of the rights holders Twapper Keeper developments will allow Twitter users to delete their tweets from the Twapper Keeper archive. In addition they will be able to opt out of the Twapper Keeper archive service.

A more challenging request has been to restrict the archiving of a Twitter user's stream of tweets to the owner. Such a requirement could hinder the development of other user requests (e.g. the ability to archive tweets from a list of Twitter users). It can also be argued that since an individual's tweets can be accessed by using the Twitter search facility users it would be unreasonable to expect that a stream of an individual's tweets should not be archived. From this perspective the issue of, for example, archiving of tweets which might be embarrassing to the poster should be regarded as an educational and new media literacy issue, on par with understanding the risks of sending inappropriate messages to public mailing lists. However we acknowledge that here is a need to address the concerns raised by this request. We are therefore planning to remove the ability to allow open archiving of an individual's tweets; instead users will need to login (via OAuth) and will only be able to create a public archive of their own Twitter stream.

It should be noted that it will still be possible to archive tweets based on keywords. Since the keywords could coincide with a Twitter ID it is possible to find tweets which may refer to an individual. Removing the ability to archive by keywords would undermine the credibility of the service and could result in users migrating to an alternative service. Our approach to this dilemma is to raise awareness of the ways in which an individual's tweets could be archived on the service's FAQ and remind users of possible risks in posting public tweets and the mechanisms for deleting tweets from the Twapper Keeper archive and from Twitter.

5.3. Sustainability Issues

As the Twapper Keeper service continues to grow various issues related to the quality of the service are beginning to arise including:

- Servers are being over-utilised.
- Continuity of backup service is not optimized.
- Users have limited visibility to items that are still being queued for archiving.
- Resource contention on various backend archiving processes needs to be tuned to support the increased number of terms being archived.

In order to address these issues the following actions have been taken:

- A dedicated server (versus a virtual server) has been procured and setup to host the service.
- Additional disk space has been added to provide primary and backup storage on the dedicated server.
- New system monitoring services have been implemented to provide status to system administrators and end users.
- On-going monitoring, refactoring and tuning of archiving algorithms in order to improve the archiving efficiency and effectiveness.

6. EXPERIENCES

6.1. Use of the Twapper Keeper Service

Twapper Keeper was used to archive tweets from the World Wide Web 2010 conference, held in Raleigh North Carolina on 28-30 April 2010 as illustrated in Figure 1. On 7 May 2010 3,616 tweets which used the #www2010 hashtag had been harvested from a total of 909 users.

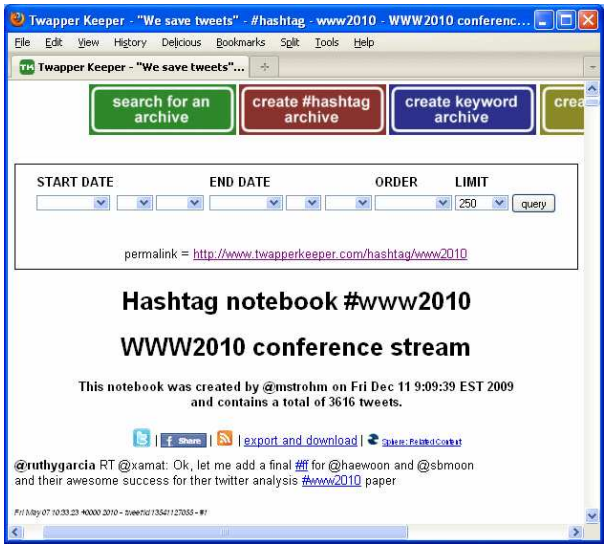


Figure 1: Twapper Keeper interface for the #www2010 archive

During the development work we became aware that tweets were missing from the archive [18]. We discovered that gaps can be introduced during disconnects / reconnects especially if there is a latency in the data transferred between Twitter and Twapper Keeper; if the latency is high, data could be lost. Twapper Keeper now runs a background process that uses the REST /search API to check to see if we have missed any tweets and then attempts to fill in the gaps.

In order to attempt to validate the coverage of the Twapper Keeper server a comparison with the WTHashtag service's archive of the #www2010 hashtag showed that, on the same date, the WTHashtag service also reported that there were 3,616 tweets.

The potential loss of tweets and possible differences in the time taken by different harvesting services to harvest tweets will be documented in a Twapper Keeper FAQ to ensure that users wishing to publicise statistics

on the numbers of tweets are aware of possible discrepancies in the figures provided by different services.

6.2. Use of Twapper Keeper APIs

The Summarizr service was developed independently of Twapper Keeper but made use of Twapper Keeper APIs. This service provides summaries and graphs of Twitter usage based on the Twapper Keeper data. This service, which was developed at Eduserv, an educational charity based in the UK, removes the need for developments having to be provided by Twapper Keeper and vindicates the decision to encourage the take-up of the APIs by others. Independent discussions are taking place with the Summarizr developer on ways in which the Summarizr statistical summaries can themselves be reused by other applications.

As well as providing statistics on the total numbers of tweets and users for a hashtag as illustrated in Figure 2 the service also displays graphs showing the top Twitterers, @reply recipients, conversations, related hashtags and URLs tweeted [21].

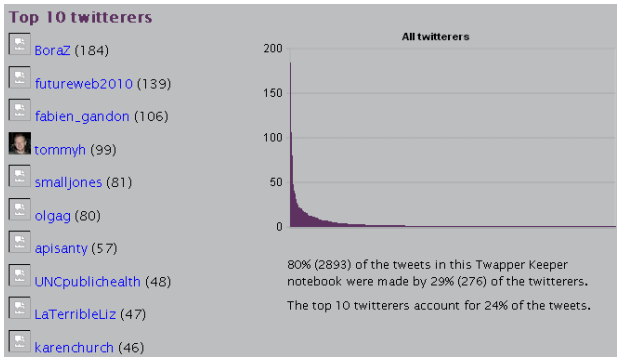


Figure 2: Use of Summarizr to display statistics of use **of** the for the #www2010 hashtag

The service also provides a display of a word cloud showing the relative frequency of the most popular words tweeted with a particular hashtag as illustrated in Figure 3.

Figure 3: Summarizr display of popular words

6.3. Summary of Status of Twapper Keeper Service Use

As of 1 July 2010 the Twapper Keeper archive contains 1,243 user archives, 1,263 keyword archives and 7,683 hashtag archives. There are a total of 321,351,085 tweets stored. The average number of tweets ingested per second is from 50 to 3,000 per minute (around 180,000 per hour. or 4.32 million per day). Since Twitter itself processes about 65 million tweets per day the Twapper Keeper service is currently processing about 6-7% of the total public traffic.

7. FURTHER DEVELOPMENTS

Recent developments to Twapper Keeper and Summarizr are storage and display of geo-location data. We invited participants at the IWMW 2010 event in July 2010 to geo-locate their tweets which enabled a map of the locations of Twitter users to be produced thus providing evidence of the remote participation at an event [12] as shown in Figure 4.

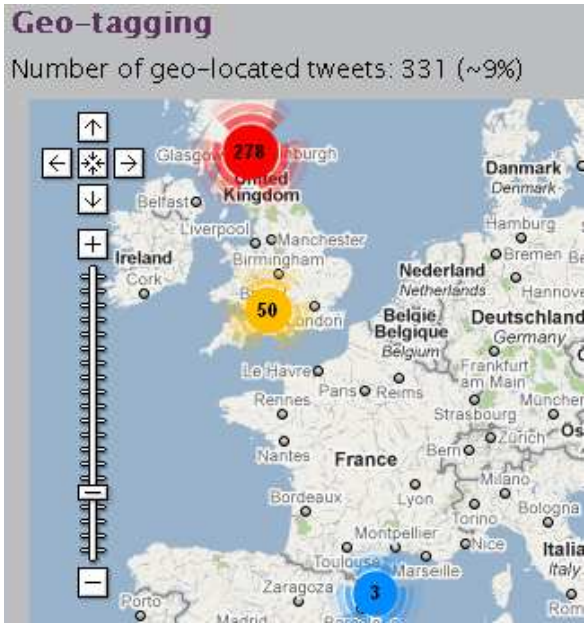


Figure 4: Display of geo-located tweets

In addition we used the Twapper Keeper APIs in conjunction with the Twitter captioning service to provide a captioned version of recordings of plenary talks shortly after the videos had been published.

We have identified the need to ensure that Twitter users are aware of the implications of Twitter archiving services. We will be developing guidelines which will help to raise awareness of the ways in which tweets could be reused, the possible risks which this may entail and approaches they can take to minimise such risks, including deletion of tweets from Twitter and archiving services which support deletion.

8. SUSTAINABILITY CHALLENGES

Although the JISC funding has been used to fund development work to address the needs of the UK higher education community and to support the migration of the service to a more stable platform this pump-priming funding cannot guarantee the sustainability of the service in the long-term.

The software developments which have been funded will be made available under an open source licence, thus allowing the Twapper Keeper service to be

recreated if the host service were to disappear. In addition the data itself is available in a rich format allowing the data to be easily migrated to other environments.

The policy decision to fund development of a service provided by a commercial provider reflects the changing funding environment in the UK's public sector, in which the government has announced significant reductions in future investments in the sector.

The approaches taken in funding Twapper Keeper developments provides a useful experiment in alternative approaches to development work which will inform other development activities funded by the JISC.

9. CONCLUSIONS

This paper has described the importance of the archiving of Twitter posts by outlining case studies based on the ability to have reliable and consistent access to tweets. Rather than commissioning a new service the JISC has funded development of an existing 'cloud' service provided by Twapper Keeper. The approaches to ensuring the sustainability of this investment have been described.

The paper has summarised the requests received from the user community on developments to the services and reviewed the technical and policy challenges which the development work has faced.

The paper has described the experiences gained in use of the Twapper Keeper service to archive tweets from a large international conference and concluded by summarising developments which were deployed at a national event in the UK.

10. ACKNOLWEDGEMENTS

Acknowledgements are given to the JISC for their support for the Twapper Keeper development project.

11. REFERENCES

[1] Diakopoulos. N. and Shamma, D. *Characterizing Debate Performance via Aggregated Twitter Sentiment.* CHI 2010, ACM, 2010.

[2] Hawksey, M. *Twitter powered subtitles for BBC iPlayer*, MASHe blog, 16 Feb 2010, <http://www.rsc-ne-scotland.org.uk/mashe/2010/02/twitter-powered-subtitles-for-bbc-iplayer/> 2010.

[3] Hawksey, M. *Gordon Brown's Building Britain's Digital Future announcement with twitter subtitles*, MASHe blog, 23 Mar 2010, <http://www.rsc-ne-scotland.org.uk/mashe/2010/03/gordon-browns-building-britains-digital-future-announcement-with-twitter-subtitles/> 2010.

[4] Hawksey, M. *Presentation: Twitter for in-class voting and more for ESTICT SIG*, MASHe blog, 20 Apr 2010, <http://www.rsc-ne-scotland.org.uk/mashe/2010/04/presentation-twitter-for-in-class-voting-and-more-for-estict-sig/ > 2010.

[5] Hirst, A. *Twitter Gardening – Pruning Unwanted Followers*, OUseful blog, 24 Sep 2009, <http://blog.ouseful.info/2009/09/24/twitter-gardening-pruning-unwanted-followers/> 2009

[6] Hirst, A. *More Thinkses Around Twitter Hashtag Networks: #JISCRI*, <http://blog.ouseful.info/2009/09/04/more-thinkses-around-twitter-hashtag-networks-jiscri/> 2009.

[7] Hirst, A. *Twitter Powered Subtitles for Conference Audio/Videos on Youtube*, OUseful blog, 8 Mar 2009, <http://blog.ouseful.info/2009/03/08/twitter-powered-subtitles-for-conference-audiovideos-on-youtube/> 2009.

[8] JISC. *JISC Strategy* 2010-2012, <http://www.jisc.ac.uk/aboutus/strategy/strategy1012.aspx> 2010.

[9] Kelly, B. *Use of Twitter at the ALTC 2009 Conference*, UK Web Focus blog, 14 Sep 2009, <http://ukwebfocus.wordpress.com/2009/09/14/use-of-twitter-at-the-altc-2009-conference/> 2009.

[10] Kelly, B. *Privatisation and Centralisation Themes at JISC 10 Conference*, UK Web Focus blog, 15 Apr 2010, <http://ukwebfocus.wordpress.com/2010/04/15/privatisation-and-centralisation-themes-at-jisc-10-conference/> 2010.

[11] Kelly, B. *Evidence on Use of Twitter for Live Blogging*, UK Web Focus blog, 4 Aug 2009, <http://ukwebfocus.wordpress.com/2009/08/04/evidence-on-use-of-twitter-for-live-blogging/>, 2009.

[12] Kelly, B. *Geo-locating Your Event Tweets*, UK Web Focus blog, 6 July 2010, <http://ukwebfocus.wordpress.com/2010/07/06/geo-locating-your-event-tweets/> 2010.

[13] Library of Congress. *How Tweet It Is!: Library Acquires Entire Twitter Archive*, Library of Congress blog, 14 Apr 2010, <http://blogs.loc.gov/loc/2010/04/how-tweet-it-is-library-acquires-entire-twitter-archive/> 2010.

[14] McGill, K. *Summary: What Is The Web?*, IWMW 2009 blog, 30 Jul 2009, <http://iwmw2009.wordpress.com/2009/07/30/summary-what-is-the-web/> 2009.

[15] Twitter. *Twitter Search API Method: search*, <http://apiwiki.twitter.com/Twitter-Search-API-Method:+search> 2010

[16] Twitter. *Terms of Service*, <http://twitter.com/tos> 2010.

[17] Twapper Keeper. *JISC-Funded Developments To Twapper Keeper*, Twapper Keeper blog, 16 Apr 2010, <http://twapperkeeper.wordpress.com/2010/04/16/jisc-funded-developments-to-twapper-keeper/> 2010.

[18] Twapper Keeper. *Study of Missed Tweets*, Twapper Keeper blog, 5 May 2010, <http://twapperkeeper.wordpress.com/2010/05/05/study-of-missed-tweets/> 2010.

[19] Twapper Keeper. *Hashtag notebook #jiscri tag for JISC Rapid Innovation projects which are creating software for Higher Education*, <http://www.twapperkeeper.com/hashtag/jiscri> 2010.

[20] Sakaki, T., Okazaki, M., and Matsuo, Y. "Earthquake Shakes Twitter Users: Real-time Event Detection by Social Sensors", *International World Wide Web Conference Proceedings of the 19th international conference on World wide web*, Raleigh, North Carolina, USA. <http://portal.acm.org/citation.cfm?id=1772690.1772777> 2010.

[21] Summarizr. *TwapperKeeper Archive for hashtag notebook www2010*, <http://summarizr.labs.eduserv.org.uk/?hashtag=www2010> 2010.

LARGE-SCALE COLLECTIONS UNDER THE MAGNIFYING GLASS: FORMAT IDENTIFICATION FOR WEB ARCHIVES

Clément Oury
Bibliothèque nationale de France
Web Archive Preservation Manager
Digital Legal Deposit

ABSTRACT

Institutions that perform web crawls in order to gather heritage collections have millions – or even billions – of files encoded in thousands of different formats about which they barely know anything. Many of these heritage institutions are members of the International Internet Preservation Consortium, whose Preservation Working Group decided to address the issues related to format identification in web archive.

Its first goal is to design an overview of the formats to be found in different types of collections (large-, small-scale…) over time. It shows that the web seems to be becoming a more standardized space. A small number of formats – frequently open – cover from 90 to 95% of web archive collections, and we can reasonably hope to find preservation strategies for them.

However, this survey is mainly built on a source – the MIME type of the file sent in the server response – that gives good statistical trends but is not fully reliable for every file. This is the reason why it appears necessary to study how to use, for web archives, identification tools developed for other kinds of digital assets.

1. BACKGROUND

Since many years, heritage institutions recognized the need to keep the memory of the material that public institutions, businesses and individuals produce and distribute thanks to the Internet. In 2003, some of them decided to group together within the International Internet Preservation Consortium (IIPC). The goals of the consortium are to collaboratively build collections of Internet content, to promote web archiving and "to foster the development and use of common tools, techniques and standards for the creation of international archives". The IIPC is currently made up of more than forty institutions. They generally use – possibly along with other techniques – crawling software, or robots, to explore the web and retrieve content that they will hold for the long term. The sets of documents harvested and produced by these robots are called web archives.

At first sight, from the point of view of formats, web archive collections may appear to be a preservation nightmare. There is no need to recall here the huge number of files harvested by crawl engines. Even the most focused archiving project has to tackle millions of files – see the Harvard University Library, whose Web Archive Collection Service dates back only from 2009 and that already has to preserve 14 million files. These figures rise to hundreds of millions of files per year for those performing crawls of entire top level domains (.au, .fr), not to mention the huge collections of Internet Archive, which in less than 15 years of existence has gathered more than 150 billion files.

The second main issue is that virtually all kind of formats are likely to be available on the Internet. At the same time, when a crawler harvests files online, it gets very little information about the formats of the documents it is capturing. The only indication generally available is the MIME type of the file that the server sends to the harvesting robot, in the http response header. Unfortunately, this information is often badly specified, peculiar (we found at the BnF a curious "application/x-something" MIME type), or even totally wrong (for example, a gif image may be indicated as text/html – webmasters do not see it as a problem for rendering, because a browser is able to read gif files directly).

In short, web archiving institutions generally have millions – or even billions – of files encoded in thousands of different formats about which they barely know anything. Heritage institutions tend therefore to turn to identification tools developed in order to ensure the preservation of other kind of digital material – or developed for other purposes than preservation.

This is the reason why the Preservation Working Group of the IIPC (or PWG) acknowledged the need to specifically address this critical issue through a dedicated work package. In this paper, we will present the goals of this work package and its methodology. We

will then look at the first outcomes, and finally present future work.

2. RELATED WORKS

Several studies have been done in order to characterize parts of the web, particularly national web domains. Their goal is to analyze the main features of the websites and web files related to a single country: notably the number of domains, the number of files per domain, the number of hyperlinks between websites[1]… In these studies, we generally find a section dedicated to formats. However, we have not identified any works specifically dedicated to file format analysis. On the other hand, there are some – even though rare – studies that examine the ability of identification tools to deal with web archives. In 2007, Bart Kiers from the Dutch National Library tested the behaviour of Jhove and Droid on web archives [5]. The test sample was limited to ten small and medium size websites, grouping 40 000 unique objects for a total uncompressed size of 2.2 Gb. Two years latter, Andrew Long from the National Library of Australia tested five format identification tools (Droid, File identifier, Jhove, TrID and the in-house developed tool Lister) on two web archive samples (from 115 000 to 18 million files) [7]. Finally, the Danish National Library and the Aarhus University Library are currently testing the use of Droid and Jhove on a 100 Tb archive [4].

3. OBJECTIVES AND ORGANIZATION

The first objective of the "format identification tools gap analysis" work package was to produce an overview of the main formats generally available in web archives (using the data obtained from a large number of institutions). It is intended to give a brief insight into the formats that were to be found on the web at different times. This is a way to participate in the general PWG goal of describing the "web technical environment" (that is what formats, software, browsers… were used on the web) over time. On the other hand, this overview should help us in comparing different collections, to identify their characteristics and their specificities.

This study is however built on information – MIME types sent in the server response – that is commonly considered unreliable. First, this has been done for practical reasons: this kind of information was the easiest to get from member institutions. Secondly, we made the assumption that even though the information was not reliable for each individual object, it was sufficient, at a larger scale, to reflect the big picture of format distribution. This assumption has been confirmed by the results of the survey. The proportions found for the only institution that used an identification tool (Library and Archives Canada, which directly ran Droid on their web archives[2]) were globally similar, from 2005 to 2009, to those we found for institutions having only sent MIME information[3].

In the survey, a first distinction is made between domain and selective crawls. Domain crawls are launched on a very large number of websites (e.g. 1,7 millions for both the .fr and .au domains in 2010), but the crawling depth is limited. Moreover, domain crawls are only performed once or twice a year. They are generally launched by national libraries in the framework of a law on digital legal deposit. On the other hand, selective crawls are performed on a more limited number of websites (from hundreds to thousands) generally chosen by librarians or archivists. Those websites may be harvested many times a year, and crawling depth is generally better.

Domain crawls are the best way to obtain a representative sample – a snapshot – of the web. According to R. Baeza-Yates *et al.* [3], crawls of national domains provide a good balance between diversity and completeness by including pages that share a common geographical, historical and cultural context but written by diverse authors in different organizations. However, even though data from selective crawls may be considered less representative (since human, subjective selection replaces automatic selection by a robot), they were taken into account because data from selective crawls may be considered as more "valuable" and may thus deserve more costly preservation actions.

It would not have been feasible to gather information for every year; so the survey focuses on arbitrarily chosen years[4]. Finally, we asked people to give only the list of the 50 most ranked formats. The ranking was calculated according to the number of objects in this format. All institutions were indeed not able to compute the number of bytes per format.

So far, we have received answers from ten institutions[5]. We can consider that this sample is representative of the diversity of the different collections IIPC members may hold: three institutions sent data for domain crawls; eight institutions sent data

[1] See for example [2] for the Danish web, [6] for the Australian web or [8] for the Portuguese web. R. Baeza-Yates et al. proposed in 2006 a comparative study of the national web characterization of several countries across the world, at various dates between 1998 and 2005 [3].

[2] Thanks to the Pronom database (http://www.nationalarchives.gov.uk/aboutapps/pronom/), we converted the Pronom identifiers into MIME types.

[3] Note as an exception a surprisingly low number of gif files in the 2005 collection (only 0.8% of the collection against an average percentage of 7%).

[4] It was decided to start from 1997 (date of the first Swedish domain crawl) and to take the years 2000, 2005, 2007 and 2009. We choose to add 2007 because more information was likely to be available for recent years (many institutions didn't start their web archiving program before 2007).

[5] Namely the national libraries of Australia (NLA), France (BnF), Netherlands (KB-NL), Sweden, the Library of Congress (LC), the British Library (BL), Harvard University Library, Library and Archives Canada (LAC), The National Archives of United Kingdom (TNA), and the Internet Archive (IA).

for selective crawls (some institutions sent data for both types of crawls). Finally, Internet Archive sent information on their crawls of the entire web.

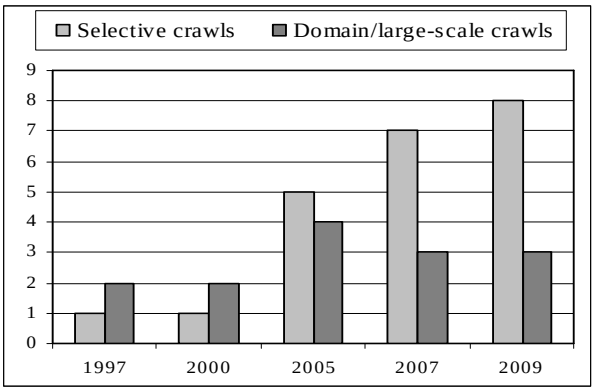


Figure 1. Types of collections in the survey.

4. FIRST OUTCOMES: GENERAL CHARACTERIZATION

4.1. Web (Archive) Trends, 1997 to 2009

As a first outcome of this study, we can draw a general overview of the main format trends in web archives. We have compiled information from Internet Archive (available from 1997 to 2005) and from domain crawls of Sweden (1997 to 2009), Australia and France[6] (both 2005 to 2009[7]). Note however that there is an unavoidable gap between the web trends and the web archive trends, because file formats that are hardly harvested by crawlers – flash files, rich media... – are under-represented in archives (and heritage institutions objective is to reduce this gap by improving their harvesting tools).

It is not surprising to see, all over the period, a strong

	1997	2000	2005	2007	2009
1	text/html	text/html	text/html	text/html	text/html
2	image/gif	image/gif	image/jpeg	image/jpeg	image/jpeg
3	image/jpeg	image/jpeg	image/gif	image/gif	image/gif
4	text/plain	text/plain	text/plain	application/pdf	application/pdf
5	application/octet-stream	unknown	application/pdf	**image/png**	**image/png**
6	application/zip	application/pdf	no-type/unknown	text/plain	text/plain
7	*application/postscript*	application/octet-stream	**image/png**	**text/css**	**text/css**
8	application/pdf	application/zip	**text/css**	app./x-javascript	app./x-javascript
9	*audio/x-wav*	audio/x-pn-realaudio	application/x-javascript	app./x-shockwave-flash	app./x-shockwave-flash
10	unknown	application/msword	app./x-shockwave-flash	no-type/unknown	**text/xml**
11	application/msword	*application/postscript*	application/octet-stream	**text/xml**	no-type/unknown
12	image/tiff	**image/png**	application/msword	**application/xml**	**application/xml**
13	application/x-tar	**text/css**	**text/xml**	application/msword	application/octet-stream
14	*video/quicktime*	audio/midi	application/zip	app./octet-stream	application/msword
15	audio/x-aiff	*audio/x-wav*	application/x-tar	image/pjpeg	application/rss+xml
16	application/rtf	application/x-tar	image/pjpeg	audio/mpeg	text/javascript
17	video/mpeg	application/x-tex	*application/postscript*	application/zip	image/pjpeg
18	app./vnd.ms-powerpoint	audio/x-pn-realaudio-plugin	audio/x-pn-realaudio	text/javascript	audio/mpeg
19	audio/x-mpeg	audio/x-midi	audio/mpeg	application/rss	application/javascript
20	Javascript	audio/x-sidtune	application/x-gzip	image/bmp	application/atom+xml
21	app./x-shockwave-flash	application/mac-binhex40	**application/xml**	image/x-icon	application/zip
22	**image/png**	image/tiff	application/vnd	app./x-zip-compressed	image/bmp
23	**application/sgml**	*video/quicktime*	app./x-zip-compressed	application/atom	app./force-download
24	**text/css**	application/x-gzip	image/bmp	application/vnd	image/x-icon
25	video/x-ms-asf	chemical/x-pdb	text/javascript	*video/quicktime*	app./vnd.ms-excel
26	x-world/x-vrml	audio/basic	image/jpg	audio/x-pn-realaudio	app./x-zip-compressed
27	application/vnd	application/vnd.ms-excel	*video/quicktime*	video/x-ms-wmv	video/x-ms-wmv
28	image/pjpeg	audio/mpeg	audio/prs.sid	*audio/x-wav*	*video/quicktime*
29	application/x-gzip	application/rtf	video/mpeg	*application/postscript*	app./vnd.ms-powerpoint
30	audio/x-midi	video/mpeg	image/tiff	app./force-download	*audio/x-wav*

Table 1. High ranked formats in large-scale collections, from 1997 to 2009 (increasing formats are in bold, decreasing in italic)

[6] Information used from the BnF for the year 2009 actually dates from November/December 2008.

[7] To compile this information, we calculated the average percentage of formats in different web archives instead of using the total number of files of each collection (e.g. if a format represents 30% of collection A, 20% of B and 10% of C, the average percentage is 20%, even though institution A holds three times more data than the two others). This principle has been applied to all computations. We did so to avoid an over-representation of big collections against smaller ones, which would have prevented all comparisons.

domination of html, jpeg and gif. It is even more impressive if we look at the percentage of files: for the year 2009, 70% of files are encoded in html, 18% in jpeg, 6% in gif. However, this chart allows us to identify the rise and fall of some formats. We may notice the destiny of png (0.006% of web collections in 1997), which now ranks in fifth place (that is… not even 1.2% of available files). Observe also the increasing rank of css and xml files (while its ancestor sgml disappeared).

On the other hand, some formats that were very popular twelve years ago now rank at a very low place. This is the case of postscript (from the 7th to the 45th place), wav audio files, and even quicktime video files. This is another surprising lesson of this overview: even though we know that the web is increasingly becoming a huge video platform, large-scale crawls don't seem able to tackle the video harvesting issue. The number of captured audiovisual files is increasing (as an example, Sweden crawled 1 300 quicktime videos in 1997, 11 000 in 2005 and 25 000 in 2009), but not as fast as the overall growth of our collections – and definitively not as fast as the percentage of audiovisual content on the web. We will see that selective crawls may provide some solutions to this problem.

From a preservation point of view, however, these figures are good news. Standardized formats are gaining ground against proprietary ones (for example jpeg against gif; xml and png are open formats).

4.2. Comparisons Between Domain Crawls

Statistically, significant differences between collections should not appear in such a mass of data. We can expect web technologies – and formats – to be equally distributed within the various countries. In fact, if we look at the collections issued from the 2009 domain crawls (France, Sweden, Australia), we find exactly the same formats in the list of the ten most ranked[8]. And we find only 36 different formats in the list of the 30 most ranked.

However, older collections do not show such strong similarities. There is a greater variety of MIME types in the list of high ranked formats for the domain crawls of previous years.

	2005	2007	2009
Top 10	12	10	10
Top 20	25	25	22
Top 30	42	39	36

Table 2. Number of different formats in the list of high ranked formats of the three domain crawls collections, 2005 to 2009.

We can thus conclude that as the web becomes more commonly used, national dissimilarities in the use of web technologies tend to fade away.

4.3. Comparing Selective and Domain crawls

A similar compilation has been made for collections issued from selective crawls. The goal was also to examine if there were significant discrepancies for collections coming from large- and small-scale harvests, and between small-scale collections from different institutions.

Again, there are no obvious differences between collections. If we compare the average distribution of formats in domain crawls with the average distribution in selective crawls, from 2005 to 2009, we notice few variations. However, a more careful analysis shows some interesting features of specific collections. At the end of the list of the 30 most ranked formats for selective crawls, we find many video formats (such as asf, windows media video or flash videos) that do not appear in domain crawls. Focusing only on formats available in large-scale collections would lead us to leave out these files.

It is also possible to identify characteristics that are related to the nature of the collection. As an example, The National Archives of the United Kingdom are entrusted with the harvesting of governmental publications and websites. This is probably the reason why we notice a larger proportion of pdf and desktop application formats[9].

Moreover, this survey allows us to discover formats that are specific to a collection, over time. For example, the proportion of flash video files which the French National Library (BnF) holds in its 2007 and 2009 selective collections is seven times higher than the average. This last case is explained by the fact that BnF launched in 2007 specific crawls of a video broadcasting platform called Dailymotion, the French equivalent of YouTube.

If we only look at major web archive trends, we will not consider Excel spreadsheets, real audio files or flash video files as being formats that deserve a specific preservation strategy. This is why institutions should also look at their own data in order to assess specific preservation needs. We should not forget the preservation operations won't apply to the web itself – they will be designed for the heritage collections derived from the web.

[8] Excluding the "no-type" format.

[9] In 2005 and 2007, twice the percentage of pdf and word files, five times the percentage of excel files.

5. FIT FOR PRESERVATION?

Following on from this, are heritage institutions familiar with such file formats? To answer this question, we can look at a report produced by the National Library of Netherlands (KB-NL). The library conducted a survey on the digital documents held by libraries, archives and museums [10]. From the replies of 76 institutions, they drew up a list of 137 different formats, of which 19 were quoted by seven or more respondents.

5 of these 19 formats only do not figure in our top 20 formats of 2009 domain or selective crawls[10]. On the other hand, the distribution is very different. For example, the most cited format in the KB-NL study is tiff (50 occurrences), while it does not even appear in the top 20 lists for web archives. Similarly, gif and html appear only at the 8th and 10th rank (against 1st and 3rd in web archives). We found similar percentages only for jpeg (2nd rank in both studies), pdf (respectively 3rd and 4th rank) and xml (4th and 8th rank).

The case of tiff files – frequently used for digitization – shows that heritage institutions producing digital documents rarely use the same formats as people that commonly publish online. Yet, can we conclude from this that web formats aren't fit for preservation? To have a first answer, let us refer to the list of "Recommended Data Formats for Preservation Purposes" established by the Florida Digital Archive [9].

Formats are classified in three categories: high, medium and low confidence level. Applying these criteria to the average distribution of 2009 selective crawls (only top 20 highest ranked formats), we can conclude that the formats available on the web are not the worst we can imagine from a preservation point of view (see table 3 below). Note that for some formats (such as html or pdf), there is a different level of confidence depending on the format version. Since this kind of information is not available in MIME type reports, we need to look at the response from Library and Archives Canada – and to assume that its sample is representative. Again, using the 2009 figures:

- xhmtl files (high confidence) represent 11% of the html files (other versions have a medium confidence grade)[11];
- on the other hand, 98% of PDF files only have a "low confidence" grade. As a matter of fact, PDF-A (high confidence) and PDF-X2 and 3 (medium confidence) respectively represent 0.5 and 1.5% of the total.

MIME Type	Average proportion	Confidence level
text/html	67,979%	High or Medium
image/jpeg	11,885%	Medium
image/gif	6,613%	Medium
unknown/no-type	3,440%	n/a
application/pdf	3,256%	High to Low
text/plain	1,286%	High
image/png	1,182%	High
text/css	0,847%	Medium
application/x-javascript	0,551%	Medium
text/xml	0,444%	High
application/x-shockwave-flash	0,326%	Low
application/atom+xml	0,187%	High
application/xml	0,180%	High
application/msword	0,167%	Low
application/octet-stream	0,114%	Medium or Low
text/javascript	0,104%	Medium
application/rss+xml	0,097%	High
audio/mpeg	0,077%	Medium
application/vnd.ms-powerpoint	0,069%	Low
application/vnd.ms-excel	0,061%	Low

Table 3. Average proportion of MIME types in 2009 selective crawls[12].

6. USING FORMAT IDENTIFICATION TOOLS WITH WEB ARCHIVES

Although this survey provides a first insight into the formats of the collections we hold, this is not enough to guarantee their preservation in the long term. First, it only gives statistical trends: at the level of each individual file, the information is not reliable. No migration operation is possible without such knowledge. Secondly, nothing is said about the format version – which stands as an obstacle for emulation strategies, because we won't emulate the same browser, say, for html 2.0 and 4.0. Therefore, whatever preservation strategy is chosen, relevancy of format information remains a critical issue.

This is the reason why the use of identification tools appears as a necessary step towards a better understanding of our collections. By identification tools, we mean all software that "describes" the format of a specific file. It can range from simple format identification to validation, feature extraction or assessment[13]. This definition may include tools such as Droid, Jhove (v1 & 2) or the National Library of New Zealand metadata extraction tool[14].

[10] TIFF, WAV, AVI, MPEG (2) and MDB files are neither in the domain nor the selective crawls list. BMP is only in the domain crawls list. XLS and PPT are only in the selective crawls list.

[11] Note that there is a specific MIME type for xhtml documents: application/xhtml+xml. However, this MIME type is very rarely used, and commonly replaced for convenience reasons by text/html. Even W3C recommends doing so. See http://www.w3.org/TR/xhtml-media-types/.

[12] Figures from 2009 domain crawls are not presented as they show very similar trends.

[13] These categories are defined in [1].

[14] Droid: http://sourceforge.net/projects/droid/
Jhove 1: http://hul.harvard.edu/jhove/
Jhove 2: https://confluence.ucop.edu/display/JHOVE2Info/Home
NLNZ metadata extraction tool: http://meta-extractor.sourceforge.net/

Previous reports have already outlined several issues that arise when using identification tools for web archives:

- Some major formats are not supported by characterization tools. For example, neither the NLNZ metadata extraction tool nor Jhove 1&2 are currently able to characterize PNG files. There is no Jhove module for MP3, even though it is the most frequent audio format within web archives…
- Files may not be well formed, which is a problem for identification. This is mainly the case for html files that are frequently hand written or modified. KB-NL reported in 2007 that none of the 20 000 processed html files were considered valid or even well-formed [5]. Let us hope that the growing use of xhtml will reduce this risk.
- Scalability and performance probably remain the major issue for web archives. Tools need to be able to process hundreds of millions of files. NLA report evaluates that it would take 42 days for Droid to identify 18 millions files (0.8 Tb) on a single-core machine, whereas up to a billion files can be harvested in few weeks during a domain crawl [7].

7. FUTURE WORK

The objectives of the PWG are now to organize a collaborative review of the main identification tools. We will build upon the format overview to organize the test protocol and to define the test samples. These tests are intended to assess the efficiency of the tools (notably by providing metrics), and report on any difficulties encountered (e.g. with specific file formats, with the management of container formats, or due to the number of files). Recommendations and best practices for using these tools will be proposed.

Finally, we hope to present a set of enhancements for these tools to address specific web archive issues and requirements. Fortunately, the institutions that are leading the development of the major tools generally hold web archives along with other digital collections, and are also IIPC members.

In addition, test outcomes will also help us to enrich the general overview of the formats in web archives. It will also be necessary to find a durable way to store, update and make available this format overview. An Excel spreadsheet was a convenient way to compile information coming from disparate sources. The work done so far can now be used as a test bench to design a real database, where each IIPC member institution could add its own data.

8. CONCLUSION

The first outcomes of this study allow us to avoid an overly pessimistic point of view: even though web archives consist of files over which we have no control, it is not impossible to ensure their preservation.

There is indeed much good news: considering the major trends, it looks like the web is becoming a more and more standardized space. Standard and open formats are gaining ground. Moreover, existing differences between "national" webs are tending to disappear. The second reassuring piece of news is that most files are encoded in a very limited number of formats. Having a preservation strategy for the ten highest ranked formats would be sufficient to render from 95 to 98% of the collection[15].

Yet, this shouldn't lead to an overly optimistic vision. The importance or the "value" of a format does not only depend on the number of files in which they are encoded. This is evident if we choose as the unit of reference not the number of object, but the size. In fact, the ten higher ranked formats (in terms of number of files) generally cover only 50 to 80% of the bytes of the collection[16]. Even the 30 most ranked formats cover only from 70 to 95% of the collection size. This is mainly due to the size of audiovisual files, which are commonly 1 000 to 10 000 times bigger than html pages. Video files may be considered by curators or researchers as more valuable – not only because they hold rich content, but also because without them, heritage web archives collections would not be representative of the "living" web. On the other hand, many html files are not "real" content, but were artificially produced by the robot, for example when it tried to extract javascript links.

Finally preservation actions need to be focused as a priority on file formats that risk becoming obsolete – and this is unlikely to be the case for the major web formats, at least in the short term. This is the reason why institutions may choose to focus on formats that they alone hold: and in this case, having an overview of what is available in other archives will be very useful. This is a way for collaborative work – at national or international level – to provide the tools, knowledge and advice to help institutions to define their own preservation objectives.

9. ACKNOWLEDGEMENTS

The author gratefully acknowledges all those who contributed to the survey. He wishes also to thank J. van der Knijff, S. van Bussel and R. Voorburg from the National Library of the Netherlands for their compilation of the existing literature on web archives formats identification.

[15] For domain crawls, from 2005 to 2009, these 10 formats are: html, jpeg, gif, pdf, plain text, png, css, javascript and shockwave-flash.

[16] Size in bytes is not available for all collections. This ratio of 50 to 80% has been computed from LC selective crawls (2005 to 2009), NLA domain crawl (2009), BnF 2009 domain crawl and 2007-2009 selective crawls.

10. REFERENCES

[1] Abrams, S., Owens E. and Cramer, T. "What? So what?': The Next-Generation JHOVE2 Architecture for Format-Aware Characterization", *Proceedings of the Fifth International Conference on Preservation of Digital Objects*, London, Great Britain, 2008.

[2] Andersen, B. *The DK-domain: in words and figures*. Netarkivet.dk, Aarhus, Denmark, 2005. Online: http://netarchive.dk/publikationer/DFrevy_english.pdf.

[3] Baeza-Yates R., Castillo C. and Efthimiadis, E., "Characterization of national web domains", *ACM Transactions on Internet Technology (TOIT)*, 2007. Online: http://www.chato.cl/research/.

[4] Jensen, C., Larsen, T., Jurik, B.O., Hansen, T.S., Blekinge, A.A., Frellesen, J.L. and Zierau, E. *Evaluation report of additional tools and strategies*, Kongelige Bibliotek, Statsbiblioteket, Aarhus, Copenhaguen, Denmark, 2009. Still unpublished.

[5] Kiers, B. *Web Archiving within the KB and some preliminary results with JHOVE and DROID*, Koninklijke Bibliotheek, The Hague, Netherlands, 2007. Online: http://www.kb.nl/hrd/dd/dd_projecten/webarchivering/documenten/IIPC-PWG-Webarchiving-JHove-DROID-test.pdf.

[6] Koerbin, P. *The Australian web domain harvests: a preliminary quantitative analysis of the archive data*, National Library of Australia, Canberra, Australia, 2008. Online: http://pandora.nla.gov.au/documents/auscrawls.pdf.

[7] Long, A. *Long-term preservation of web archives – experimenting with emulation and migration methodologies*, National Library of Australia, IIPC, 2009. Online: http://www.netpreserve.org/publications/NLA_2009_IIPC_Report.pdf.

[8] Miranda, J. and Gomes, D. "Trends in Web Characteristics", *Proceedings of the 2009 Latin American Web Congress*, Washington, United States of America, 2009. Online: http://arquivo-web.fccn.pt/about-the-archive/trends-in-web-characteristics.

[9] "Recommended Data Formats for Preservation Purposes in the Florida Digital Archive". Florida Center for Library Automation, United States of America, 2008. Online: http://www.fcla.edu/digitalArchive/pdfs/recFormats.pdf.

[10] Van Bussel, S. Ad Houtman, F. *Gap analysis: a survey of PA tool provision*, Koninklijke Bibliotheek, The Hague, Netherlands, 2009. Online: http://www.planets-project.eu/docs/reports/PA2D3gapanalysis.pdf.

Session 9a: Building Systems

FULL PAPER

A DATA-FIRST PRESERVATION STRATEGY: DATA MANAGEMENT IN SPAR

Louise Fauduet
Bibliothèque nationale de France
Department of Preservation and Conservation

Sébastien Peyrard
Bibliothèque nationale de France
Department of Bibliographic and Digital Information

ABSTRACT

The Bibliothèque nationale de France has developed its trusted digital repository, SPAR (Scalable Preservation and Archiving Repository), as a data-first system. This implies having fully described collections, through use of metadata standards in the information packages, such as METS, PREMIS, MIX or textMD, in a way that will make sense given the diversity of our documents.

The need for full documentation also applies to the system itself. On the one hand, SPAR is self-describing in order to ensure its durability. On the other hand, all the information that is ingested into the system contributes to determine its settings and its behavior. The Data Management module is at the heart of these information flows.

We expect to push this data-first objective ahead by using RDF technology, based on existing and trusted information models and ontologies, such as OAIS and PREMIS. The challenges and successes we encounter all serve the greater goal of having a unique and versatile data model for every user of the system, whether collection curator or system manager.

1. INTRODUCTION

The SPAR system, Bibliothèque nationale de France's trusted digital repository, is finally stepping out of the design phases and becoming a concrete tool in preservation and collection management at the BnF (See Bermès and al. [1]).

SPAR is conceived as a data-first system, where data is used both to curate the collections and to manage the system.

The collections are fully described and each piece of information should be individually accessible. The flexibility in querying information is intended to make collection management as easy as possible from a preservation perspective.

The system is fully self-describing: every process is documented within it; and it can be set up by the ingested data, without having to change the actual implementation of the system.

The data-first approach is a three-part endeavor. First, it depends on the way the OAIS information model is implemented in the information packages. Then, it relies on the translation of the OAIS functional model into SPAR's architecture, making the Data Management module possible. Last but not least, the use of RDF enables BnF staff to draw on the data in order to manage the system and the collections.

2. DESCRIBING THE COLLECTIONS: OUR METADATA STANDARDS

2.1. The Metadata Makings of an AIP

2.1.1. METS: the Why

Each digital document is ingested into the SPAR preservation system as an Information package, as defined by the OAIS model, with a METS manifest as packaging information stored within each package. Expressing our information needs in a standardized way and in compliance with best practices facilitates maintenance and is therefore a great ally in digital preservation.

METS, like other preservation metadata formats, offers great flexibility, and many further choices are required in order to implement it — which sections to use, which other metadata formats to embed, which granularity levels to define in order to describe the package, and so on.

The challenge of these numerous implementation choices prompted librarians to reflect on best practices which would fit the BnF's specific needs without reducing interoperability[1], even if actual exchange between repositories is not in our short- or medium-term plans. One of the greatest advantages of METS is indeed its wide use in the digital preservation world in general and in libraries in particular. Its active user community facilitates METS's implementation while protecting against format obsolescence.

[1] On this issue, see for instance Rebecca S. Guenther [4].

2.1.2. METS: the How

The abstract quality and great genericity of OAIS along with the flexibility and openness of METS made the implementation of both in the BnF context a great step in itself. The main choices that had an impact on the coverage of the collections by the metadata, involve METS sections, granularity levels, and embedded information.

First, we chose to exclude from our METS implementation the metsHdr, structLink and behaviorSec, for which we had no need, and the rightsMD subsection, since we would rather have a dynamic calculation of the legal status of a document at the time it is accessed (See Martin [5]).

The main factor in the choice of granularity levels in METS's structural map was the great diversity of material to be ingested in SPAR: digitized texts and still images at first, then digitized and born-digital audiovisual content, Web archives, the library's born-digital archives, and so on. The adoption of generic terms to describe the levels within the digital object avoids the heavy maintenance of a specific vocabulary.

Therefore, four levels were adopted in the structural map. From the broader to the narrower, they are:

- set: ensemble of groups. This level is only intended to contextualize groups by describing a higher level, which is purely intellectual. E.g. serial, or multivolume monograph.
- group: the reference level in our repository. It is the level at which a digital document is digitized and/or manipulated. E.g. physical volume of a monograph; CD…
- object: an intellectual subdivision of a package E.g. page of a document, side of a vinyl…
- file: a concrete file.

Regarding embedded schemes in the dmdSec and amdSec sections of METS, three main decisions were made.

Dublin Core is implemented in dmdSec and sourceMD: using METS from a preservation perspective, we don't need to include in AIPs the type of highly structured descriptive information that already exists in our catalog[2]. This type of information shows what the package is about, but is independent of the actual digital embodiment of the document; it is not needed to make preservation plans. More pragmatically, its non inclusion in the packages avoids close dependencies and mutual updates between two systems, our catalog and SPAR, so that the Archive is as autonomous as possible.

However, some specific information needs, expressed by SPAR's users at the librarian end, require more elements than the DC's 15 standard ones: description of the institution detaining the files requires Qualified DC; domain specific identifiers such as ISSN, ISBN, bar code, call numbers or even pagination types required more specific elements that did not exist as such in DC; so we used our own schema, adding as few elements as we could. This infringement on our interoperability vow is a compromise that enables a better management of librarian needs.

Finally, we use premis:object and premis:event in the techMD and digiprovMD sections of METS, because of PREMIS' genericity and closeness to the OAIS, and of the wide adoption of the "METS + PREMIS" duo among libraries.

However, premis:object is not intended to express text-, image-, sound- and video-specific file characteristics. To this end, we use the METS-proof and widely adopted MIX scheme for image files and textMD for text files.

An overall consensus on a characterization format for audio, and above all video content, has yet to be reached in the digital preservation community. Few schemas are able to express every piece of information our audiovisual experts need for collections management in a well-structured and thus easily manageable form. Conversely, few are designed to be used inside packaging information, and thus make elements we express in other sections of METS mandatory.

Our double need of expressivity and modularity brought us to MPEG-7, an ISO standard, suited to both audio and video, and even to multimedia and program files. Therefore we rejected more widely adopted standards for audio files, such as AES-X098B.

2.2. Describing a Preservation System with Data: Reference Information Packages

The choices we made regarding METS define our SIPs, AIPs and DIPs in a way that satisfies our information needs as to the digital documents we preserve. Yet there is an equally important type of information that also has to be preserved in SPAR: all the documentation regarding the way the system works and the nature of the information that is preserved in it. In order for SPAR to be self-referenced and OAIS-compliant, this information is enclosed in information packages as well, in a category that we named reference information packages. They can be of three different types: context, formats and agents.

Context reference information allows us to create links between ensembles of packages that share certain characteristics. In SPAR, this mainly means assigning packages to their relevant track and channel. A track is a family of documents with similar intellectual and legal characteristics: there is digitized printed content track, a Web legal deposit track, and so on. Each track is divided into channels, which share homogeneous technical

[2] http://catalogue.bnf.fr/

characteristics[3]. Description of each channel and track is factorized in a dedicated information package. In the future, we intend to use information packages to describe software environment in an emulation perspective.

We also give representation information about every format for which we have designed a preservation strategy. This can include standards such as TIFF 6.0, or BnF profiles restraining these formats, for instance uncompressed 24 bits TIFF in 300 dpi resolution.

Finally, SPAR ingests reference information about agents performing preservation operations, which can be human (administrator, preservation expert), software tools (identification, characterization and validation tools) and processes in SPAR (such as the ingest and package update process).

Grouping information that is common to many digital objects is just one feature of reference packages. They have maintenance enhancement advantages: updating this central information means it is not necessary to update every information package that relates to it.

They also materialize a genuine "data-prior-to-system" approach: these information packages allow us to set system parameters with machine actionable files. For instance, the system can check the conformity of image files with a specific profile of TIFF used at BnF (TIFF 6.0, 24 bits, 300 dpi resolution, BnF watermarking, etc.) each time a package with files whose MIME type is identified as image/tiff is ingested. In this way, data defines and configures processes, not the other way around. This enhances control of the system processes by users that are not IT specialists.

Last but not least, the reference information packages include a sample file or the source code of the tool, with human readable documentation about the format, in order to meet the needs of digital curators and preservation experts. Every aspect of the system functionalities that has an impact on librarianship is documented in SPAR.

3. WORKING THE DATA INTO THE SYSTEM: THE DATA-MANAGEMENT MODULE IN SPAR'S ARCHITECTURE

Having defined the types of data that go into SPAR, we will examine how they are processed and used by the system — to the extent that certain types of ingested data actually determine the settings of the system.

[3] For instance, the channel B of the Audiovisual track contains the product of the digitization of analog audio and video document acquired through legal deposit, with well-described and easily manageable production formats; whereas the channel A of the same track concerns legal deposit of born digital content (excluding documents harvested on the web), which we are constrained to ingest "as is", with inevitably unknown or misused formats.

3.1. A Modular Implementation of the OAIS

From its early stages of inception, SPAR was to be a modular system: in order to allow easier integration of new technology, each main function had to be able to be improved at its own pace. Thus the system was divided into modules following the OAIS functional model entities: Ingest, Data Management, Archival Storage, Access, Administration, and Preservation Planning, the last one to be developed at a later date. They form SPAR's "core".

Additional modules which do not have a direct equivalent in the OAIS functional model have been designed, such as a Rights Management module, which is not yet implemented, or Pre-Ingest modules for each specific ensemble of similar material. The Pre-Ingest phase is meant to harmonize the different digital documents into a SIP that is SPAR-compliant and can be processed in the rest of the system in a generic way.

In this environment, Data Management could be considered as the inner sanctum of the system, along with Storage. It centralizes all the existing data in the system according to a unified data model, making it accessible through the same interface. See Figure 1 below.

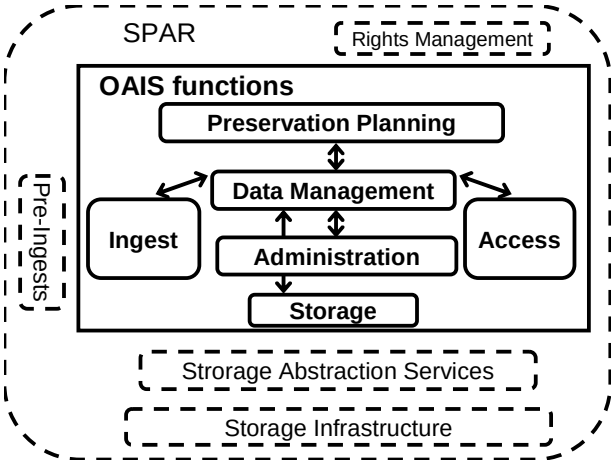


Figure 1. Data Management within SPAR's modular architecture.

3.2. The Data Management Module as SPAR's Information Hub

3.2.1. Data Flows Between Modules

The Data Management module, or DM, is not directly accessible by a human user: every interaction with it is mediated by another module, be it Ingest, Storage, Access or Administration.

All of these interactions have been designed from use cases developed during the specification stage of SPAR. Most of the use cases involve more than two modules, but DM's role in all of them can be viewed as an information hub, managing the metadata flow. All these

interactions use RESTful Web services technologies that are compliant with our modular design.

Data Management intervenes in two stages of the ingest process: first during the creation of a SIP, when the latter's characteristics are audited to check their conformity with the channel requirements, stored in the Data Management module, then at the end of the ingest process when the metadata contained in an AIP are recorded into DM.

The Storage module interacts with DM to query reference data and make sure storage requirements for an AIP are met.

Information exchange between the Data Management and Access modules is maybe the most important one for curators and IT staff to achieve their collection management goals, since any retrieval of data for use out of the system is mediated by Access, whether the data is simply identifiers or more structured information about the system or the packages. Access also needs information from DM in order to provide DIPs to the users.

Data Management's abilities to sift and reorder information are naturally used by the Administration entity in the daily toil of the system, and should assist the future Preservation Planning in preparing migrations and other preservation actions.

3.2.2. Setting Parameters With Data

Data Management's role as a central nervous system of SPAR can be illustrated with the example of one particular type of data: the Service Level Agreements (SLAs) contained in channel reference packages.

As seen in paragraph 2.2, channels are defined for a particular set of homogeneous digital material which requires the same services from the Archive. The producers of these digital documents and the Archive write down the exact nature of their commitments to one another in a human-readable agreement, which is transcribed in a machine-actionable set of SLAs, written in XML according to an in-house schema. The exact equivalence of human- and machine-readable SLAs guarantees the user communities that the services agreed upon with the Archive are actually implemented as such. These SLAs, along with schematrons to validate the specific METS profiles used in the channel, form a channel reference package.

For each channel, there are three SLAs: one for ingest, one for preservation and one for access issues. Indeed, the same type of controls, such as file format or number of copies, may be applied very differently in the varying stages of the ingestion / preservation / dissemination process. For instance, for the same package, the SIP and DIP may be stored only once, while the AIP will be stored in several copies.

The SLAs define four types of requirements. Requirements at the channel level include the SLA's validity dates, the opening and closing hours or the maximum unavailability duration of the system, for instance. There are also requirements on packages (minimum and maximum size of package, allowed and denied format types for the channel, AIP retention duration, and so on), on storage (number of copies, presence of encryption, etc.) and on processes, determining how the system's resources can be mobilized by the channel (minimum and maximum number of invocations of a process for a given period and so on). All those requirements are entered into the Data Management module when a channel reference package is ingested, and set system variables.

To see how this data is used in the daily workings of SPAR, and the Data Management module's role in them, we can take the "Ingest a SIP" use case as an example.

Whenever the Ingest module receives notification of a new SIP, it is audited, and its METS manifest is validated using reference data that has been put into DM, notably information from the channel package: which users are authorized to submit packages in this channel, or what the METS profile for the SIPs of this channel is.

Then, using DM's capabilities as an index of all the packages in SPAR, the system checks the SIP's identifiers against those of the AIPs already stored to determine whether the SIP is a brand new package or an update, and if so, what type of update.

The SIP's characteristics are checked against the channel service level agreements in DM, such as the maximum size or the number of objects allowed in the package.

The files are individually identified, characterized and validated using tools documented in DM through reference packages. The result is compared with the list of formats accepted in the channel, listed in the SLAs. The behavior of the system if the criteria of the SLAs are not met (rejection of the package or mere warning to administrators) is also specified in the SLAs stored in DM.

Finally, a unique identifier is created for the package, and all the new metadata are added to the package's METS manifest, before the AIP is stored in the Storage module. At the same time, the information present in the METS file is added to the Data Management module.

3.3. The Inner Workings of the DM Module

3.3.1. Different Repositories for Different Needs

The Data Management module as a whole is a data repository, but it is actually divided into a Reference documents repository and a Metadata repository. The Reference documents repository contains documents used in controlling the validity of data and metadata, such as XML schemas and schematrons. The Metadata repository contains representation information and preservation description information that has been

transformed from its submitted XML encoding into RDF/XML when inserted into the Data Management module.

The choice of RDF triple stores was made following an extensive risk analysis based on the desired features of the main metadata repositories in SPAR (see 4.1.1). Resource Description Framework has a very generic and versatile data model, where the information is expressed in triples, following the syntax subject/predicate/object. It came ahead in the analysis due to its very flexible query language, SPARQL, and its good performances in mapping from the existing XML metadata and in reversibility. The benefits and challenges of that choice will be further examined in part 4.

The Metadata repository is actually composed of three separate RDF repositories. Metadata from all the AIPs in SPAR goes into the Complete metadata repository, where it is available for complex queries by the digital collection curators. From the complete metadata, a lighter, faster Selected metadata repository is extracted, to fulfil the metadata needs of the modules of SPAR themselves. Additionally, all the content of the reference packages, which is heavily used in the workings of the system, has its own Reference data repository. See Figure 2 below.

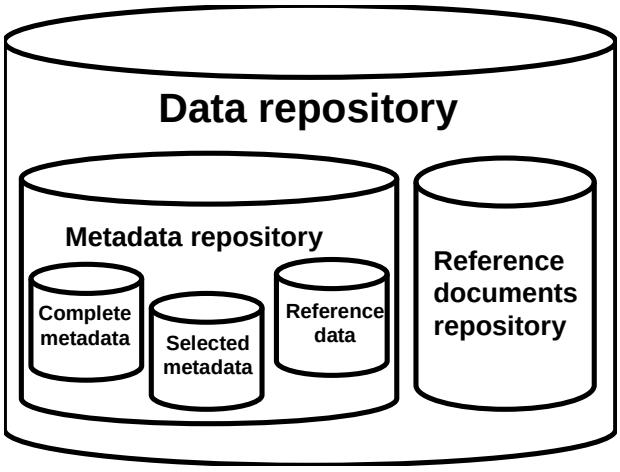


Figure 2. Data repositories in the Data Management module.

3.3.2. Making Changes in the Data Model Possible

In order to be useful, data repositories have to be up-to-date. Mechanisms are implemented in order to reconstruct the metadata repositories when new packages are added and updated. However, given the amount of metadata and reference information in the SPAR system, we had to accept compromises, and devise fail-safes.

The Complete metadata repository is not an exact one-to-one transposition of each metadata entry in the METS files of each package: some of the information is not expressed; some of it has been aggregated. For instance, the format of each individual file is not expressed in the triple store; instead the types of format a fileGrp contains will be listed for each fileGrp.

The choice of what information to keep in the RDF triple stores was based on a clear principle: it should be information that the system's users may need to query in order to select and retrieve packages according to an identified professional use. Once the packages have been retrieved and accessed via RDF requests, more detailed actions can be taken after examining the METS files themselves. Detailed examples are provided in 4.2.2.

Of course, some of the information we need in order to identify certain AIPs may have been overlooked in our initial METS to RDF mapping, and our activities will probably change over time (See Bermès and Poupeau [3]). Moreover, the data model may evolve to include new types of information we hadn't foreseen. Thus, the METS files are archived independently, and may be the basis of a planned reconstruction of the Complete metadata repository.

4. THE RDF DATA MODEL: HOW TO SPEAK THE SPAR LANGUAGE

4.1. Principles and Methodology

The risk analysis that was performed when the Data Management module was designed pointed to RDF triple stores as the least risky choice of four, when compared to relational databases, XML databases and search engines. Three families of risks were evaluated:

- risks in setting up the technology in SPAR, which included integrating the technology into the system's modules, and mapping the data from METS to the chosen solution;
- risks in managing the metadata: RDF scored very well in querying capabilities, but had higher risks regarding update features;
- risks in maintaining the technology over time: RDF's handling of data models was a plus, but the technology was still new at the time (see 4.2.2 and 4.2.3).

The choice of RDF in itself is far from enough to build an efficient data model. No domain specific ontology, that is, RDF vocabulary, existed in digital preservation when we started building the data model, so we had to build it from scratch according to the following principles.

4.1.1. Using the OAIS Information Model

While building our RDF data model, we had the same guidelines as when implementing METS: genericity, interoperability, therefore better maintenance and durability.

Since RDF aims at describing things in a self-declarative fashion, using RDF requires the

implementation of a domain specific terminology. In order to structure our own information model, we naturally turned to the OAIS information model, which was at an abstract level, thus generic, and had a very strictly standardized, documented and hierarchized terminology of concepts, which favored interoperability.

We built an ontology per OAIS information type: representation, structure, fixity, provenance and context. An additional class was built, agent, since it was a very well-identified domain by itself. It related as much to context information as to provenance information, and matched an existing PREMIS entity.

4.1.2. Reusing Existing Ontologies

One of the great features of RDF is its modularity: parts of existing ontologies, such as properties and classes, can be integrated into other ontologies. In reusing those parts that are already well-modeled and widely used, SPAR's data model gains a better conformity to existing standards, and we gained time to concentrate on developing our specific classes and properties. However, we are also bound by the intentions of these other ontologies' creators and should not bend these existing rules.

We reused Dublin Core properties[4] for descriptive information, in our reference ontology; OAI-ORE[5] and its concept of aggregation in our structure ontology, to describe relationships between granularity levels; FOAF[6] for agent information and more specifically DOAP[7] for software agents; and so on.

4.1.3. Naming Resources with URIs: info:bnf and ARK.

In RDF, resources and properties must be named with URIs. BnF already implements the ARK (Archival Resource Key) URI scheme for its digital material and metadata records. Its open source, non-proprietary nature and maintenance by a public institution (California Digital Library) made it an ideal scheme to use in a digital preservation context as well.

ARK is particularly suited to identify concrete objects, since it can point to parts or specific views of the document with "qualifiers"[8]. For instance,

- ark:/12148/bpt6k70861t names a AIP containing a digitized edition of Charles Baudelaire's 1857 *Les Fleurs du Mal*;
- ark:/12148/bpt6k70658c.version0 names the initial version of this digital document;
- ark:/12148/bpt6k70658c/f5.version0 names the 5th page of this document;
- ark:/12148/bpt6k70658c/f5/master.version0 and ark:/12148/bpt6k70658c/f5/ocr.version0 respectively name the image and ocr files for this page.

Thus, ARK is the way we name actual AIPs, or parts or them, to say something about them in RDF.

But ARK is not suitable for naming abstract information in SPAR, that is, specific properties and classes of our ontologies. ARK names have to be opaque whereas the self-declarative philosophy of the Semantic Web, and usability issues of course, require significant URIs.

To this purpose, SPAR uses the info:uri scheme. For instance info:bnf/spar/provenance# is the URI naming the representation ontology in the system, and info:bnf/spar/provenance#digitization names the abstract event "digitization".

4.2. The Result: Ontologies and Access to Data

4.2.1. An Ontology: Provenance

The provenance ontology in SPAR is very close to the PREMIS data model and shows many features of RDF, as Figure 3 below demonstrates.

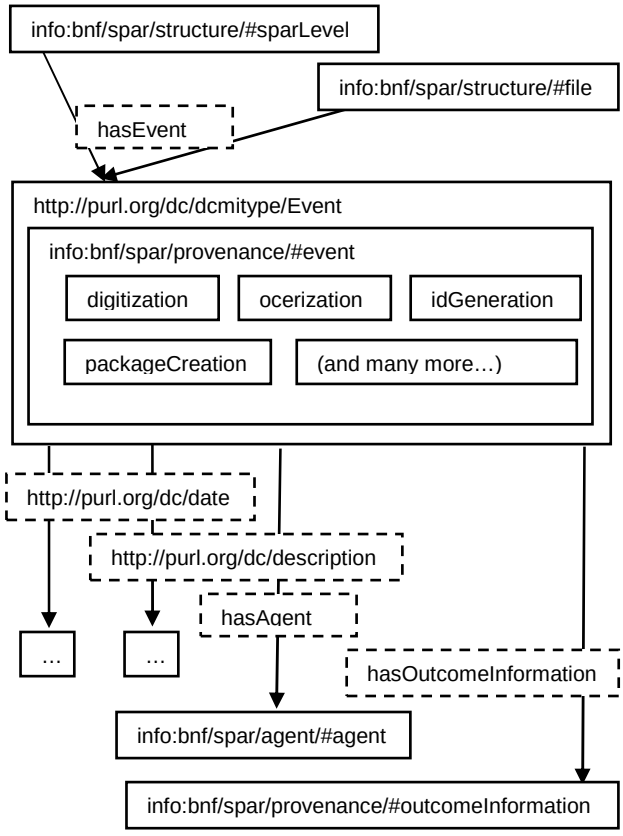


Figure 3. A simplified view of the provenance ontology

As in PREMIS, each event is viewed as an entity relating on one hand to an object, which can be at

[4] http://www.purl.org/dc/elements/1.1/ for simple Dublin Core and http://www.purl.org/dc/terms/ for qualified Dublin Core.
[5] http://www.openarchives.org/ore/1.0/rdfxml/
[6] http://xmlns.com/foaf/spec/
[7] https://usefulinc.com/doap/
[8] That is, suffixes beginning with "." or "/".

various granularity levels, and initiated on the other hand by an agent, be it human or software.

Each particular <premis:eventType> in SPAR's METS implementation in XML is modeled as a distinct class with "event" as a common superclass in RDF. For example, the digitization eventType becomes the class info:bnf/spar/provenance#digitization, being a subclass of info:bnf/spar/provenance#event.

Existing properties are reused to express some premis:event elements, as http://purl.org/dc/elements/1.1/description for eventDetail, viewed as the description of an event, or http://purl.org/dc/elements/1.1/date for eventDateTime.

4.2.2. Access to Data

The advantages of RDF listed above are particularly valuable when it comes to data retrieval issues.

Data is controlled, thus access is controlled: the same concepts and things always have the same name, that is the same URIs. Queries are precise because they go through controlled access points. And, contrary to what happens with relational databases technologies, it is not necessary to know the names of the categories of data in advance to formulate a query: they can be deduced from the way the data is structured, by successive queries.

Moreover, RDF's query language, SPARQL, is independent of the way the data is actually written down: the Data Management module uses RDF/XML, but the queries use the abstract way the data are modeled in the subject/predicate/object fashion. Although SPARQL has its own set of rules, compared to other query languages, it follows a common language pattern and is thus more intuitive. And simple query sentences can be assembled to create complex queries.

Here are some examples of queries we can formulate about material from the digitized books and still images collections:

- Which package has pages flagged as containing a table of contents, but no table of contents file in XML, which would allow dynamic navigation in the document? Answering this question helps plan retrospective creation of structured tables of contents.
- How many packages were ingested in SPAR the last month, with their number of files, the formats and the quality rate of the OCR? This traditional question shows that data also helps administrators monitor the system.
- Which packages in our digitization channel have invalid HTML table of content files? Invalid HTML doesn't necessarily impede access to the document, but is certainly harder to preserve; such a query helps preservation experts plan invalid HTML files regeneration.

4.3. Challenges and Uncertainties

Even though the BnF sees many advantages in the use of RDF to manage the data in its digital trusted repository, there are many uncertainties and problems attached to adopting a relatively new technology, mainly performance, maintainability and training issues.

4.3.1. Too Much Information?

RDF remains a recent technology with the weaknesses inherent to its newness, which we faced when implementing Data Management. First, compared to other technologies, few software providers are available for RDF triple stores; only Virtuoso suited our needs in terms of data volume and performance, and yet its implementation required a great amount of tuning and optimization. Its performances are also slower for the moment than those of traditional relational databases. Even though it may not be a foremost issue in a preservation perspective, quick response times give valuable comfort to digital curators.

This problem is exacerbated by one of the principles presiding to SPAR's creation: to use as many open source programs as possible, in order to reduce specific developments, benefit from other communities' maintenance, and enhance financial viability.

However, tests conducted in 2008 showed that our implementation of a Virtuoso Open Source triple store reached its limits when the data volume nears 2 billion triples — although it should be noted that the performances of RDF technologies are improving steadily. 2 billions may seem like a high maximum, but, considering the first channel of documents to be ingested in SPAR already includes 1 million packages with an average of 200 files and at least 5 types of metadata expressed in METS at file level[9], this amounts to 1 billion triples for basic file-level information in one single channel.

Hence the distinction between information useful to identify and access the packages, which is indexed in RDF, and information only needed once the digital documents are retrieved mentioned in 3.3.2. It enabled us to reduce considerably the amount of data indexed in the Data Management module the first channel to enter SPAR in order to gain computing power, while maintaining usability.

4.3.2. New Technologies, New skills

Using RDF had other immediate drawbacks for the staff of the BnF, be it on the IT or on the librarian side.

On the IT side, Semantic Web technologies were previously unused at BnF, and require training, first for the digital preservation team, then for their collaborators. Day-to-day monitoring of the Data

[9] That is, the MIME type of the file, its size, checksum, checksum type, and the information that each file is a file.

Management module is also more difficult, since there is little peer support or experience feedback yet.

On the librarian side, training issues are even greater, since SPAR, as a digital collection preservation and management tool, is not only intended to be used by digital preservation experts, but also by producers of data-objects and collection curators (see Bermès and Fauduet [2]). They have to understand SPAR's data model in order to express their information needs. Digital preservation experts and digital data producers may have to act as an intermediate in the beginning, but ideally, everyone dealing with digital collections should be able to get the information they need directly from Data Management, which implies learning how to query it with SPARQL.

Moreover, the lack of well-established best practices in RDF modeling for digital preservation forced us to build SPAR's data model and the ontologies "on the fly", using common sense and professional experience in data modeling.

But all these are difficulties in the short or medium term. In a long-term perspective, RDF has real organizational advantages, as it allows the separation of technical/IT issues from data/librarian ones. As complex as RDF and SPARQL can seem to be in the beginning (but is MARC any easier?), they give librarians a better control of their data, which also equates, in a data-first approach, to a better control of the system processes.

Ultimately, we hope that SPAR's data model, and its use of RDF technologies, will allow all BnF's staff dealing with digital collections preservation and curation to speak a common language that will adapt to different missions and different time constraints.

Every person in interaction with the Archive will have to refer to the same data model, using the same request language, whether they are planning long-term preservation actions such as migrations; have short-term decisions to make, requesting a new ocerization on certain documents for instance; or need the day's latest statistics. And eventually, all these users will have to define the necessary evolutions of the data model together. This could be the best way to integrate SPAR into the large and diverse ecosystem of the Bibliothèque nationale de France's activities; data-first, the rest should follow.

5. REFERENCES

[1] Bermès, E et al. "Digital preservation at the National Library of France: a technical and organizational overview", *World Library And Information Congress: 74th IFLA General Conference And Council*, 2008. Online at http://archive.ifla.org/IV/ifla74/papers/084-Bermes_Carbone_Ledoux_Lupovici-en.pdf [last consultation 2010-05-04].

[2] Bermès, E. and Fauduet, L. "The Human Face of Digital Preservation: Organizational and Staff Challenges, and Initiatives at the Bibliothèque nationale de France", *Proceedings of iPRES 2009: the Sixth International Conference on Preservation of Digital Objects*, 2009. Online at http://escholarship.org/uc/item/6bt4v3zs [last consultation 2010-04-20].

[3] Bermès, E and Poupeau, G. "Semantic Web technologies for digital preservation: the SPAR project", *Proceedings of the Poster and Demonstration Session at the 7th International Semantic Web Conference (ISWC2008)*, 2008. Online at http://sunsite.informatik.rwth-aachen.de/Publications/CEUR-WS/Vol-401/iswc2008pd_submission_14.pdf [last consultation 2010-05-04].

[4] Guenther, R. "Battle of the Buzzwords: Flexibility vs. Interoperability When Implementing PREMIS in METS", *D-LIB Magazine*, July/August 2008. Online at http://www.dlib.org/dlib/july08/guenther/07guenther.html [last consultation 2010-04-20].

[5] Martin, F. "Dynamic management of digital rights for long-term preservation: the expert system approach", *Proceedings of iPRES 2004: the Fourth International Conference on Preservation of Digital Objects*. Available online at http://ipres.las.ac.cn/pdf/Martin_presentation_Martin.pdf [last consultation 2010-05-04].

FULL PAPER

DEVELOPING INFRASTRUCTURAL SOFTWARE FOR PRESERVATION: REFLECTIONS OF LESSONS LEARNED DEVELOPING THE PLANETS TESTBED

Brian Aitken
University of Glasgow
Humanities Advanced Technology and Information Institute

Matthew Barr
University of Glasgow
Humanities Advanced Technology and Information Institute

Andrew Lindley
Austrian Institute of Technology
Safety and Security Department
Vienna

Seamus Ross
University of Glasgow
Humanities Advanced Technology and Information Institute
and
iSchool at
University of Toronto

ABSTRACT

The Planets Testbed, a key outcome of the EC co-funded Planets project, is a web based application that provides a controlled environment where users can perform experiments on a variety of preservation tools using sample data and a standardised yet configurable experiment methodology. Development of the Testbed required the close participation of many geographically and strategically disparate organisations throughout the four-year duration of the project, and this paper aims to reflect on a number of key lessons that were learned whilst developing software for digital preservation experimentation. In addition to giving an overview of the Testbed and its evolution, this paper describes the iterative development process that was adopted, presents a set of key challenges faced when developing preservation software in a distributed manner, and offers a real-world example of how lessons can be learned from these challenges.

1. INTRODUCTION

Planets (Preservation and Long-term Access through NETworked Services)[1] was a four year project, partially funded by the European Community, that ran from 2006 until 2010. Its primary goal was to build practical services and tools to help ensure long-term access to digital cultural and scientific assets [7]. The sixteen consortium members brought together and further developed a huge knowledge base of digital preservation research, with expertise pulled from national libraries, archives, leading research universities and technology companies.

Planets developed software that addressed several aspects of the digital preservation challenge. A variety of preservation action services were released to actively aid in the process of the preservation of data. These include services for migrating data, such as the SIARD suite of tools for migrating relational databases to XML [8], and services for presenting data in emulated environments, such as GRATE [14]. Planets also focussed on the development of characterisation services which could extract properties from data and perform automated comparison of such properties, and the XCL [3] Extractor and Comparator were the principal outcomes of the project in this respect. A further aspect of digital preservation that the project addressed was the need for preservation planning services that can assess an organisation's specific preservation requirements and capabilities to help define a suitable preservation plan. The Plato [2] application was developed for this purpose.

In addition, Planets also identified the need for a Testbed for digital preservation experimentation, a collaborative research environment where preservation tools and services could be systematically tested and empirical evidence on their effectiveness and applicability could be gathered, analysed and shared. The need for such a research environment can be traced back to two related projects, the Dutch Digital Preservation Testbed project [12] and the DELOS Testbed for Digital Preservation Experiments [6]. These studies identified the need for research into digital preservation to be more engineering focussed, with a clearly defined rationale and methodology and an emphasis on a controlled set of experimentation to provide justification and validity to the choice of preservation approaches and services. Planets significantly developed and refined the underlying principles of these earlier projects, resulting in the web-based Testbed application that is now available to all interested parties for preservation experimentation.

[1] http://www.planets-project.eu

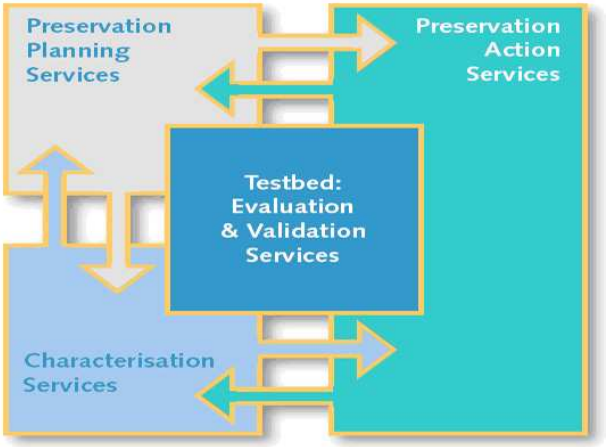


Figure 1: The Planets Software Components

Through the Testbed's online interface[2] the outputs of Planets are made available for experimentation, from preservation action and characterisation services through to the executable preservation plans generated by the Plato application. Figure 1 demonstrates how the other software outputs of Planets interact with the Testbed application. The overall aim of the Testbed is not merely restricted to validating the success of Planets-developed software; the remit of the Testbed is considerably broader and a wide variety of third party preservation focussed tools are also made accessible for experimentation.

The overall aim of the Testbed is not merely restricted to validating the success of Planets-developed software; the remit of the Testbed is considerably broader and a wide variety of third party preservation focussed tools are also made accessible for experimentation.

The background to the Testbed and an overall description of the facilities it has to offer has already been published in a number of papers [11, 1]. The primary focus of the current paper is to firstly give a general overview of the final version of the Testbed that was released during the Planets project, and then to investigate more closely the issues involved when engaging in a distributed preservation software development project. As the domain of digital preservation matures it is likely that an increasing number of preservation tools and services will be developed, both by research projects and by commercial organisations. By presenting and analysing some of the issues encountered during the development of the Testbed it is hoped that future development projects can learn from these issues and be prepared for certain challenges that are likely to emerge during the development process.

Figure 2. The Planets Testbed version 1.2

2. OVERVIEW OF THE FINAL VERSION OF THE TESTBED

The final version of the Testbed that was released during the Planets project was unveiled in April 2010, and a screenshot of this version can be viewed in Figure 2. The culmination of four years of development through eight point releases and several sub-point releases, this version of the Testbed provides a solid base for preservation experimentation through an easy to use web-based interface. To enable experimentation on preservation tools, access to these tools must be provided through the controlled experimentation environment. Within the final version of the Testbed roughly fifty preservation tools are available, each of which can be executed by an experimenter using nothing more than a web browser and an internet connection. Each preservation tool is published in the Testbed via a web-service wrapper which exposes certain aspects of a tool's functionality, specifically those aspects that have particular relevance for preservation tasks. This 'networked services' approach is a core principal of the Planets project and it offers a standardised means of accessing preservation tools, providing users with the ability to execute experiments on tools that have a disparate set of hardware and software requirements, all from a standardised web-based access point.

Preservation tools which are wrapped as services and deployed in the Testbed are split into different categories depending on their function, thus enabling experiments of different focus to be designed and executed. Services offered include migration services, such as OpenOffice, Gimp and SIARD, characterisation services, such as the New Zealand Metadata Extractor and the XCDL Extractor and emulation services (identified within the Testbed as 'CreateView' services), including Qemu and GRATE. Other service types, such

[2] https://testbed.planets-project.eu/testbed

as identification and validation, are also offered and a complete list can be found through the Testbed website.

Access to sample data is also critical to successful preservation experimentation within the Testbed. The Testbed enables users to define a dataset for their experiment in three ways: by providing their own data, by accessing the Testbed corpora of sample data, or by combining these two approaches. Access to several corpora of test files are made available to experimenters through the Testbed interface. These corpora, comprising over eleven gigabytes of files, have been collected by the Testbed team during the course of the project and provide a broad range of test files that cover not only the major office, image, sound and video formats but specific versions of such formats where applicable. In addition, the corpora include a variety of 'edge case' files, such as GIF files that have experienced bit-rot. To ensure corpora files are ideally suited for experimentation, the properties of each file are documented using XCDL, with these measurements being stored alongside the files within the corpora.

In order to test the effectiveness of preservation tools the Testbed provides facilities to measure and analyse properties relating both to the tools and the digital objects that are manipulated by tools during experimentation. Property analysis represents the principal manner in which preservation tools are evaluated in the Testbed. Properties relating to a tool include its execution time and the success of its invocation while properties of digital objects include a very broad range of properties that can vary depending on the file type and the file contents. Example digital object properties include file size, bit depth, character encoding and sample rate.

The Testbed offers a variety of services that can automatically extract and measure properties for particular file formats, including the XCL tools, the New Zealand metadata extractor, Droid and Jhove. In addition, properties can be manually measured and the Testbed provides a predefined selection of properties plus facilities enabling experimenters to define new properties. By comparing the properties of the original digital objects with the post-preservation action digital objects, and taking into consideration properties of the preservation action tool during execution, it is possible to gather a detailed understanding of the effectiveness and suitability of the tool in question. The final version of the Testbed also provides facilities to evaluate individual property measurements, thus making it more straightforward to pinpoint strengths and weaknesses encountered during an experiment.

3. EVOLUTION OF THE TESTBED

As previously mentioned, the notion of a Testbed for digital preservation experimentation had its roots in the Dutch and DELOS Testbeds. From these relatively modest beginnings Planets aimed to significantly expand the capabilities of a digital preservation Testbed, providing web-based access to experiments, online experiment execution and a shift in emphasis to the automation of tasks such as experiment execution and property measurement. These core aims of the Testbed remained relatively static over the four-year duration of the project, but the details shifted markedly as understanding of the concepts grew and knowledge of the capabilities and limitations of the architecture developed.

Rather than leaping blindly into one single, lengthy and chaotic development period, the Testbed team followed the principals of iterative software development, with an initial period of detailed requirements capture feeding into a prototype, which in turn was tested, with feedback leading to a refinement of certain requirements that were then the target of a subsequent release. This process was repeated several times, with each release resulting in a greater level of functionality and a better understanding of the underlying requirements, which may have evolved significantly since the initial period of requirements elicitation. The iterative approach adopted by the Testbed developers was the Rational Unified Process (RUP) [10], and Figure 3 demonstrates how the incremental releases of the Testbed fit into the four phases and six disciplines of RUP.

At the beginning of the project, members of the Testbed group engaged in a period of requirements elicitation. This involved several face-to-face meetings where members of the team met and discussed the goals of the project and their role within it. This included a hands-on session with the Dutch Testbed software and the involvement of representatives from other strands of the Planets project in order to ensure that their notions of a Testbed were represented during the critical phase of requirements definition. This period lasted roughly six months and during this time documents were created that helped refine the initial direction of Testbed development. This began with a set of interviews with the content holding partners within the project, which gathered information on the facilities and functionality each partner desired from a Testbed environment, for example one partner defined a scenario involving the upload of a dataset, the passing of this data through a characterisation service, then through a migration service and finally through another characterisation service in order to compare the input and output results.

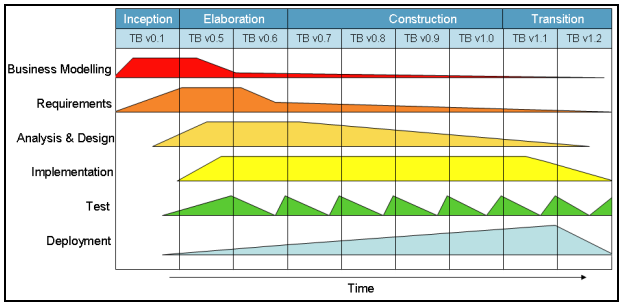


Figure 3: Development of the Testbed versions within RUP

From the interviews a series of user scenarios were formulated, representing a distillation of the core functionality required by the various partners. From the user scenarios a further abstracted set of use cases and potential actors was defined, with each use case consisting of such items as ID, title, actors, preconditions and success scenarios. Roughly 60 use cases were defined for such tasks as uploading data to the Testbed and defining experiments. The next step in the Testbed design process was the creation of functional and non-functional requirements documents, which deconstructed the information contained within the use cases into short, demonstrable statements covering every aspect of intended functionality. The requirements document followed an industry standard template [9], with each requirement being assigned a unique ID, a priority level and references back to the originating use cases. The document could be referenced by members of the Testbed group and the wider Planets project to gain an understanding of the feature-set the developers hoped to be able to develop.

The Testbed requirements document defined what the developers aimed to achieve during the course of development. However, it was not the intention at this stage to define exactly how these requirements should be implemented. The final stage in the initial design phase was the construction of a software design document, where formal definitions of the software components of which the Testbed would comprise were first formulated, class diagrams were mapped out, initial mock-ups of the Testbed front-end were proposed and the intended development environment and pre-existing software implementations were decided upon.

The initial detailed design phase of the Testbed lasted roughly six months, and by the end of this period a comprehensive set of requirements and design documents had been created, discussed, and refined. Following on from this the developers spent a further six months on the initial development of the Testbed API and the Testbed front-end, resulting in Testbed version 0.1, an HTML mock-up of the main pages of the Testbed that exhibited no real functionality but represented with a fair degree of accuracy the overall structure and layout of the final Testbed product.

Over the course of the remaining three years of the project eight major Testbed point releases were made, each of which expanded upon and refined the functionality found in the previous release. The implementation period for each release was between four and six months in duration and for each point release an implementation plan was formulated. Each implementation plan expanded upon the initial design documentation based on an increased understanding of the field, the capabilities of the software, feedback and requests from other project partners, and feedback from more formal testing sessions arranged by other members of the Testbed team.

The domain of digital preservation is not static; new research is constantly being published and the Testbed facilities which content holders desired and considered to be of the highest importance changed markedly over the course of the Planets project. Where possible a face-to-face meeting of all involved parties was held prior to the formulation of an implementation plan to ensure that feedback from the previous release could be gathered, areas where a divergence of understanding between developers and content holders could be pinpointed and addressed and the focus of the implementation period could be defined. The relatively short implementation periods and focussed point releases enabled the Testbed developers to address specific issues in each release and by publishing all implementation plans, minutes and supporting documentation on the project wiki it was the developers' intention to ensure that the decision making process and development status were as transparent as possible. As Testbed development progressed the iterations became gradually shorter and more focussed, taking on many characteristics from an agile software development framework such as Scrum [4], where a small focussed development team prioritises requirements and adapts to changing requirements through regular team meetings and updates.

4. CHALLENGES ENCOUNTERED AND LESSONS LEARNED

Throughout the four year development period of the Testbed the team noted some specific challenges and difficulties, some of which are unique to the domain of digital preservation, others which are more generally applicable to distributed software development projects. Each of these challenges has been a learning process and in the majority of cases the team identified a means to meet each challenge, or learned how to better address a similar situation in future. In this section a selection of these challenges and the lessons learned are presented.

4.1. Developing a Preservation System for a Variety of Stakeholders is Difficult

Planets involved a variety of different organisations, including national libraries and archives, research universities and technology companies. Different types of organisation and even different organisations of the same type had dissimilar and at times conflicting requirements for and demands from the Testbed software. Reaching a consensus as to the direction of development when 16 partner organisations are involved is difficult. From a logistical point of view it is infeasible to gather representatives from all organisations in one physical or even virtual location with any degree of frequency and even if such a gathering can be managed it is difficult for agreement to be reached.

This difficulty may be further exacerbated by a number of factors, as experience from the Testbed can demonstrate. Firstly, as a research project Planets involved many researchers from an academic background. Active and at times heated debate is crucial to the formulation of new ideas and to defend existing points of view, especially when researchers from different backgrounds interact. What is deemed less critical for such researchers is to reach a consensus on each discussion point, yet for software developers a conclusion to debates and a very definite pathway to follow is hugely important. Secondly, different representatives from partners organisations may be present at different meetings, and there is no guarantee that each partner institution will have a shared internal vision of the importance of certain aspects of the preservation software. Thirdly, the opinions of the stakeholders are not static; they evolve and change over time. Features that a stakeholder may consider of the utmost importance in year one of a project may easily become of minor consequence by the fourth year.

The Testbed team had to contend with these issues over the course of the project. Due to the conflicting nature of some requirements it was impossible to please everybody. For example, some partners deemed it of critical importance that certain Testbed experiments could be performed 'in private', with no experiment data being shared with other experimenters, thus enabling users to practice with the Testbed without exposing their mistakes or sensitive data to others. Conversely, other partners considered it vital that all experiments should be shared with other users in order to build up the knowledgebase, the concern being that if users were given the option of experimenting in private then few experiments would be made publicly available and some experiments that ended in failure, but which still contained ground-breaking findings, would be hidden from view.

In order to address these issues the Testbed team attempted to find a middle ground that suited a majority of stakeholders where possible. As alluded to earlier, each implementation period featured a phase of internal testing where Planets partners could give their feedback on the current iteration, and building this feedback loop into the development period helped to minimise the risk of partners having unrealistic expectations of the software. The dissemination amongst partners of all plans and minutes also helped to alleviate this issue, and the iterative design method that was adopted ensured that requirements and overall goals were fluid enough to deal with a shift in focus over time.

4.2. Distributed Development is More of a Challenge than Development at a Single Location

When engaging in the development of a preservation system, especially within the context of a research project where development is frequently entering into unknown territory, having developers working in isolation at different locations is not the ideal situation. Although there are many online collaborative tools that can help alleviate this issue, nothing is as effective as sharing an office with other developers and having the option of bouncing ideas back and forth.

The principal developers of the Testbed were based at three different organisations in three countries. In order to ensure effective communication the developers conducted weekly conference calls where open issues of a technical, design or organisational nature could be discussed and solutions could be formulated. In addition to this, the developers made frequent and efficient use of instant messaging systems to keep in contact and the use of a Subversion code repository ensured that code developments could be regularly distributed to other developers while minimising the possibility of conflicts within the code.

Effective use of such online collaborative tools was crucial to the successful operation of a distributed software development team, yet regular face-to-face meetings still proved to be essential. These helped to bolster the relationships between the developers leading to a stronger and more unified group, they improved developer morale and motivation and they also proved vital to problem-solving and decision-making. Having a day-long face-to-face developer meeting every few months provided a significant boost to productivity and was absolutely critical to the success of the Testbed, and on average three such meeting took place each year of the project. In addition to this the Testbed developers engaged in occasional longer 'exchange' visits, where a developer from one organisation travelled to and worked at another organisation for several days. These visits also proved to be highliy valuable to the development of the software.

4.3. Preservation Software Development Can Require a Significant Outlay of Developer Effort

Estimating resources for a software development project is a tricky business. This problem is not limited to the

development of preservation software or to distributed software development, but it must be taken into consideration when a project is being planned. If a project plan detailing workpackages, effort and timescales must be created and agreed upon before the official launch of a project and if project-specific requirements elicitation and systems design tasks cannot commence until after an initial plan has been compiled it is unlikely that any initial software development timescales will be accurate.

The difficulty of estimating required effort was encountered within Planets with respect to the Testbed. In the initial plan it was assumed that the Testbed would be released within the first 18 months of the project, and that this release would be stable, fully tested, documented and usable by both project partners and external parties. This estimate proved to be unrealistic, which had an impact on a range of other project activities that had been planned. In retrospect, the reaction to delays in the release of the Testbed was perhaps not as prudent as it could have been. Workpackages and deliverables that relied upon a fully operational Testbed were not redesigned to take into consideration the updated circumstances and this led to some parts of the project being less effective than they otherwise might have been.

During the course of Planets new versions of the project plan were compiled every 18 months and to a certain extent the need for more developer effort for the Testbed and the need for more realistic timescales were reflected. However, developer effort proved to be a continuing point of difficulty for the Testbed throughout the project. Overall developer effort assigned to the Testbed as an average throughout the project was less than two full-time equivalents, and this was generally split between several individuals who were working part time for the Testbed. The final release of the Testbed demonstrates just what is possible to achieve with such a limited amount of developer effort but future projects should recognise that software development does require a significant amount of developer effort, and that a degree of flexibility must be built into timescales, deliverables and follow-up activities.

4.4. Staff Turnover Will be an Issue for a Project with a Multi-Year Duration

A project that lasts four years and involves sixteen organisations cannot possibly expect to maintain the same staff for the duration of the project. It is inevitable that staff will move on and new members will join. This can have both positive and negative impacts on the project. New members can bring new ideas and innovative ways of looking at previously established practices and concepts, however there is also the risk that staff who leave do not pass on their knowledge and expertise, and that the project is unable to find suitable replacements.

During the development of the Testbed both positive and negative aspects relating to staff turnover were encountered. Within the first 18 months of the project two Testbed members left, resulting in a period of several months where the involvement of certain partners was ambiguous. Thankfully another project partner offered to provide effort for Testbed development and the supplied member of staff proved to be extremely beneficial to both the development of the application and the refinement of the core Testbed concepts. The existence of an extensive body of documentation about the Testbed, both in terms of design documentation and wiki-based plans and definitions was crucial for ensuring new staff members could gain a detailed understanding of the Testbed in the shortest possible time.

A further staff related issue that must be considered is the potential difficulty in attracting people with a suitable skill-set, especially if a project is part-way through its lifespan. The Testbed required developers with detailed practical experience of JavaEE[3], the Java Server Faces web application framework[4] and the JBoss application server[5] and finding candidates with such expertise who were willing to work on a relatively short-term research project proved to be a challenge. During the final year of Testbed development a key developer was promoted within his organisation, which would have resulted in the end of his involvement with the Testbed. The organisation in question advertised for a suitable replacement to take over development responsibilities but was unable to find anyone who was considered appropriate. The organisation allowed the existing developer to continue his involvement with the Testbed on a part time basis, but this illustrates the difficulties that a potential project must take into consideration with regards to staff turnover.

4.5. When Developing Preservation Software it is Crucial that the End Product is Developed with Long-Term Access in Mind

When developing software it is imperative that the functional and non-functional requirements of the intended users are identified. Within the context of digital preservation it is vital that in addition to this the long-term access requirements are also taken into consideration. Digital preservation practitioners extol the benefits of adhering to software standards, utilising open, non-proprietary software and formats where appropriate and ensuring adequate documentation is recorded. Software developed for digital preservation must lead by example in this respect.

The Testbed, and indeed the majority of the software developed during the Planets project took these concerns

[3] http://java.sun.com/javaee/
[4] http://java.sun.com/javaee/javaserverfaces/
[5] http://www.jboss.org/

into consideration. The Testbed was developed using the widely available and platform independent JavaEE and Metro technology stacks, with the widely established MySQL[6] database used for experiment data storage. The Testbed code is stored in a Subversion repository and has been released under an Apache2 license. It is possible for anyone to download, inspect and further develop the code from the Planets Sourceforge site[7].

However, some problems were encountered with the underlying technology used by the Testbed during its development. Due to the requirements of the core functionality provided by the Planets Interoperability Framework, the Testbed was reliant on a very specific version of the JBoss application server for the majority of the development period. This in turn required any computer on which the Testbed was compiled to be running an out of date version of Java, with newer versions causing errors. This reliance on an outdated version of Java was identified as a potential problem and was addressed during the final project year, illustrating the need to keep up to date with software developments whilst ensuring backwards compatibility with older software versions.

4.6. There Can be Conflicts and Dependencies Between Different Parts of a Large-Scale Preservation Software Development Project

If a project is large enough to be developing more than one piece of software through individual software teams then care must be taken to ensure that any interdependencies between these pieces of software are well documented and that delays or difficulties encountered by one team have a minimal effect on other teams. If one piece of software requires the delivery of a component being developed by another part of the project then effective communication between the teams is required and contingency plans that ought to be followed in the result of delays should be specified. Also, if different software applications are being developed within a project care must be taken to ensure that there is a clear distinction between the applications and that duplication of effort is kept to a minimum.

The Testbed is one part of a suite of software that was developed by the Planets project, with other software development taking place concurrently, including infrastructural software that falls under the banner of the 'Interoperability Framework' (IF), preservation tools, and other online applications such Plato.

The IF team was responsible for developing the core functionality required by the Planets applications, such as data and service registries, single sign-on services, and the workflow execution engine. Each of these components was required by the Testbed yet IF development was undertaken simultaneously with Testbed development. In some respects this approach was very valuable; Testbed and IF developers collaborated closely and the requirements of the Testbed were well reflected in the IF output. However, problems were also encountered when IF developments took longer than anticipated. In some instances the Testbed was unable to meet its deadlines due to unavoidable delays with the release of IF software, and in other cases the Testbed group had to create and rely upon mock-up functionality for the short term. Close collaboration between the two groups ensured that such delays were communicated as swiftly as possible but difficulties were still encountered when certain events such as formal testing sessions had already been scheduled. A more effective approach may have been to ensure that the core functionality provided by the IF was already available for use before the development of the Planets applications commenced.

Within Planets there was also a certain degree of conflict between two of the applications being developed, namely the Testbed and Plato. Both applications shared a common origin, specifically the Testbed work carried out by DELOS. Under the umbrella of the Planets project a divergence of aims took place, with Plato focussing specifically on the generation and evaluation of organisation-specific preservation plans and the Testbed focussing on the benchmarking of specific technical capabilities of preservation tools under certain conditions within a controlled environment. Towards the beginning of the project the Testbed and Plato teams worked on their applications without a great deal of interaction and midway through the project it was observed that a certain degree of convergence had occurred, leading to some uncertainty and conflict between the two teams. Having identified the risk of convergence a greater effort was made to define clear boundaries between the two applications, a strategy that proved to be successful. From this point onwards the two teams engaged more closely and shared ideas and code more frequently, reducing any duplication of effort and ensuring both applications were interoperable where appropriate, specifically with results aggregation from the Testbed feeding into Plato and executable preservation plans from Plato being testable within the Testbed environment.

The Testbed group identified the lack of an overall software architect within the Planets project and would recommend such a role in a future project. The principal benefits of a software architect are twofold. Firstly s/he would be in a position to form an overall picture of the software developments and to a certain extent shape these developments and ensure that each independent development group is both aware of the work of others and can be presented with a distilled vision of where the work of their group is placed within the broader canvas of the project. Secondly s/he would be able to act as a buffer zone between the blue-sky research undertaken by

[6] http://www.mysql.com/

[7] https://sourceforge.net/projects/planets-suite/

academics and the software developers, who require very definite and clear plans for development.

4.7. Effective communication is a challenge within a large-scale project

In addition to communication challenges relating to a distributed development team as mentioned above, it was observed during the course of the project that communication between different workpackages and project areas was at times difficult to manage. With so many partners involved and such a wide variety of research and development activities being undertaken people tended to focus on their own silo rather than being able to formulate a complete picture of the project. This is a very difficult challenge to overcome in such a large project. The sheer number of publications, deliverables, wiki pages, and meetings means that simply keeping up to date with developments in one project area takes considerable time, and following the outputs of the entire project is much less feasible. This can result in synergies between different groups being missed and increases the risk of duplication of effort.

One area of Planets where this problem was effectively addressed relates to digital object properties. As mentioned earlier, these are vital to the evaluation of tool performance within the Testbed and for a long time different parts of the project were engaging in research into digital object properties independently and without much collaboration or awareness of each other's work. Towards the middle of the project members of the Testbed group became aware that a gap between different parts of the project needed to be bridged and moved to define a Planets-wide digital object properties working group. This working group brought a variety of project strands together and resulted in a shared Planets conceptual framework for digital object properties within the context of digital preservation, leading to some valuable research outcomes [5] and a standardised ontology based approach to properties that was adopted by the project as a whole.

5. CONCLUSION

Over the course of the four years of the Planets project the Testbed group successfully followed an iterative development approach to design, develop and refine a web-based application that both fulfilled the original remit and met the additional needs that were identified during the project. The end product is a stable and feature-rich web-based environment that can serve as a very solid base for research and experimentation within the field of digital preservation. The experiments database provides an extremely useful knowledgebase of the performance of digital preservation tools than can help broaden the understanding of digital preservation issues, and further experimentation can be continued through the application itself.

By the end of the Planets project more than one hundred external users had signed up as Testbed experimenters, with access to the Testbed's online presence being provided by HATII at the University of Glasgow. Active research into preservation using the Testbed has been carried out by Planets partners, for example one study performed research on the migration of a large corpus of TIFF images while another study investigated emulation, virtualisation and binary translation. External users have also begun using the Testbed to pursue their own research, and by the end of the project the Testbed environment had begun to receive positive online reviews [13].

As this paper has demonstrated, developing preservation software presents a number of challenges, especially when many disparate stakeholders are involved and the project duration spans many years. These challenges may be organisational in nature, such as issues relating to a distributed development team and the danger of conflicts and dependencies between development groups. They may relate specifically to staffing, such as the difficulty of managing staff turnover and attracting new staff with the correct skill-set. Challenges may also be of a technical nature, such as ensuring the software being developed follows best practice in digital preservation and ensuring a suitable development process is pursued. The Testbed team has addressed these challenges and has produced a stable product that can be further built upon and developed by subsequent projects.

Although Planets ended in May 2010, the Open Planets Foundation[8] (OPF) has since been established to continue the innovative and highly beneficial digital preservation research and development that was spearheaded by Planets. The Testbed will continue to be managed, developed and supported by the OPF for the foreseeable future.

6. ACKNOWLEDGEMENTS

Work presented in this paper was supported in part by the European Union under the 6th Framework programme through intially the DELOS NoE on digital libraries (IST-507618), and then mainly through the Planets project (IST-033789).

7. REFERENCES

[1] Aitken, B, Helwig, P, Jackson, A, Lindley, A, Nicchiarelli, E and Ross, S, "The Planets Testbed: Science for Digital Preservation" *in. Code4Lib, vol. 1*, no. 5, June 2008. [Online]. Available: http://journal.code4lib.org/articles/83

[8] http://www.openplanetsfoundation.org/

[2] [2] Becker, C, Kulovits H, Rauber, A and Hofman, H. "Plato: a service oriented decision support system for preservation planning," *JCDL '08: Proceedings of the 8th ACM/IEEE-CS joint conference on Digital libraries*, New York, NY, USA: ACM, 2008, pp. 367–370, [Online]. Available: http://doi.acm.org/10.1145/1378889.1378954

[3] Becker, C, Rauber, A, Heydegger, V, Schnasse, J and Thaller, M. "A generic xml language for characterising objects to support digital preservation" in, *SAC '08: Proceedings of the 2008 ACM symposium on Applied computing*., New York, NY, USA: ACM, 2008, pp. 402–406.

[4] Cohn, M. Succeeding with Agile: Software Development Using Scrum, Addison Wesley, Boston, 2009.

[5] Dappert, A .and Farquhar, A. "Significance Is in the Eye of the Stakeholder " in *Research and Advanced Technology for Digital Libraries*, Springer, Berlin, 2009, pp. 297-308. [Online]. Available: http://www.springerlink.com/content/r473n317t030/

[6] DELOS, "“DELOS deliverable WP6, D6.1.1, Framework for Testbed for digital preservation experiments" in. 2004. [Online]. Available: http://www.dpc.delos.info/private/output/DELOS_WP 6_D611_finalv2

[7] Farquhar, A. and Hockx-Yu, H. ”Planets: Integrated Services for Digital Preservation”, *The International Journal of Digital Curation,Issue 2*, Volume 2, pp 88-99, 2007. [Online] Available: http://www.ijdc.net/index.php/ijdc/article/view/45

[8] Heuscher, S, Jaermann, S, Keller-Marxer, P and Moehle, F. "Providing authentic long-term archival access to complex relational data," in Proceedings PV-2004: Ensuring the Long-Term Preservation and Adding Value to the Scientific and Technical Data, 5-7 October 2004, European Space Agency. Noordwijk, 2004, pp. pp. 241–261. [Online]. Available: http://arxiv.org/abs/cs/0408054

[9] IEEE IEEE Recommended practice for software requirements specifications, IEEE Std 830-1998, 1998.

[10] Kroll, P and Kruchten, P. The Rational Unified Process Made Easy: A Practitioners Guide to the RUP, Addison Wesley, Boston, 2003.

[11] Lindley, A, Jackson, A and Aitken, B, "A Collaborative Research Environment for Digital Preservation - the Planets Testbed" in, *1st International Workshop on Collaboration tools for Preservation of Environment and Cultural Heritage at IEEE WETICE 2010*, [Online]. Available:http://planets-project.ait.ac.at/publications/PlanetsTestbed_COPECH _08032010.pdf

[12] Potter, M. "Researching Long Term Digital Preservation Approaches in the Dutch Digital Preservation Testbed (Testbed Digitale Bewaring)," in *RLG DigiNews*, Vol 6 No 3, [Online]. Available: http://digitalarchive.oclc.org/da/ViewObjectMain.jsp?fileid=0000070519:000006287741&reqid=3550 #fea ture2

[13] Prom, C. Planets Testbed Review Practical E-Records Blog, 2010 [Online]. Available: http://e-records.chrisprom.com/? p=1183

[14] von Suchodoletz, D., van der Hoeven, J. ”Emulation: From Digital Artefact to Remotely Rendered Environments”, *The International Journal of Digital Curation*, Issue 3, Volume 4, pp146-155, 2009. [Online] Available: http://www.ijdc.net/index.php/ijdc/article/view/141

BUILDING BLOCKS FOR THE NEW KB E-DEPOT

Hilde van Wijngaarden **Judith Rog** **Peter Marijnen**

Koninklijke Bibliotheek
Prins Willem-Alexanderhof 5
2595 BE The Hague
The Netherlands

ABSTRACT[1]

The National Library of the Netherlands (KB) will renew its digital archiving environment. The current system, the e-Depot with DIAS by IBM as its technical core, has been operational since 2003 and needs to be updated. More importantly, a new system is required because KB has published a new strategic plan with ambitious goals. They require development of an infrastructure that can process, store, preserve and retrieve millions of digital objects, now and for the long term. The digital collections will include e-journals, e-books, websites and digitized master images and will grow from 20 TB currently to 720 TB in 2013. The New e-Depot will also implement tools for digital preservation, as being developed in international collaboration (Planets, JHOVE, etc.).

Together with eight European national libraries, KB defined the architectural framework for the new system. It is based on a modular approach and translated into so-called building blocks for a preservation environment. This paper discusses the building blocks and the rationale for the components-based architecture of the New e-Depot. Currently, requirements for all the building blocks are finalised. A market consultation for the workflow component will starts in the summer of 2010 and the procurement process for the other components will follow in the fall. The first iteration of the New e-Depot will be delivered in 2012.

1. RENEWING THE E-DEPOT

In January 2010, the new strategic plan of the Koninklijke Bibliotheek, the National Library of the Netherlands (KB), was published [1]. This new strategic plan is an ambitious plan with a strong focus on the digital library: digitisation, online access and long-term storage. To put this plan into action, KB needs an infrastructure that can process, store, preserve and retrieve millions of digital objects, now and for the long term. The current digital processing and archiving environment, the e-Depot, cannot fully address the new challenges and will be replaced by a new, improved and extended processing and long-term preservation environment.

Digital archiving and permanent access has been a key priority of the KB since the late nineties of the 20th century. After experiments and prototyping, KB and IBM developed an archiving environment between 2000-2002. In March 2003 the current e-Depot, with the IBM system DIAS [3] as its technical core was taken into production. Since then, more than 15 million e-journal articles from major international publishers have been loaded into the system.

This environment will be renewed for the following reasons:

- KB sets out to process and preserve multiple types of digital collections while the current environment is tailor-made for processing and managing e-journal articles.
- KB needs to upscale its processing and storage environment for:
  - processing at least ten times as many digital items in a limited time frame as it does currently;
  - processing digital items that will be much larger then they are currently;
  - storing and managing at least twenty times as many Terabytes than it does currently (estimation: up to 720 TB in 2013).
- Functionality for identification, characterisation, format-conversion, and other newly developed preservation functionality has to be added to the system to ensure permanent access.
- Software combinations that are used in DIAS have reached their 'end-of-life'. Although all components are standard IBM products and are still supported, their current combination in DIAS is becoming vulnerable.
- The KB-IBM maintenance contract will expire in September 2012.

[1] This paper reflects the work and writings of the New e-Depot team at KB, consisting of Judith Rog, Jeffrey van der Hoeven, Aad Lampers, Yola Park, Peter Marijnen, Liedewij Lamers and Maarten van Schie. This paper is a joint paper of the whole group.

First plans to renew the e-Depot environment have started in 2007. This included a collaborative effort to set requirements for digital preservation functionalities and services with the Deutsche National Bibliothek (DNB) and the Niedersächsische Stats- und Universitätsbibliothek Göttingen (SUB). During the period March to October 2009 this international collaboration was extended and renamed to the LTP Working Group. Several meetings were held with representatives of eight National Libraries in Europe (Spain, Portugal, Switzerland, Germany, UK, Czech Republic, Norway and the Netherlands) to explore the possible collaboration in developing and implementing a next generation long-term preservation system. Together, the libraries worked on scoping and defined a modular approach and so-called building blocks for a preservation environment. To actually enter into a Request for Information (RfI) process together turned out to be too challenging due to different needs and planning- and budget constraints. However, cooperation was continued on further information sharing and working towards common long-term preservation services [4]. The eight national libraries decided to include each other in their development/procurement processes with sharing information and if possible inviting each other to join in meetings with suppliers.

2. SCOPING THE LONG-TERM PRESERVATION SYSTEM

One of the outcomes of the international working group was what we called the 'two-layered OAIS-model'. When starting to work on joint requirements, we started with a discussion on scope. The OAIS-model describes the necessary depot-functionalities for a long-term digital archive. But what does this mean when translated to detailed requirements? How much does a long-term digital archive have to 'do' when compared to the wider digital library infrastructure, or even the library functions as a whole? In our view, a library consists of a number of depots and the OAIS-functions are applicable to each of them. This starting point opens the need to define which functionality should be realised at library-level and which functionality should be realised at (e-) Depot-level. The previously mentioned LTP Working Group agreed to a two-layered OAIS model-approach as presented in the picture below. It defined which (part(s) of) OAIS-functions should be centralised at library level (i.e. identity management, billing-functionality) and which (part(s) of) OAIS-functions are executed at depot-level. This resulted in the picture shown in figure 1. During requirements elicitation for the New e-Depot, this model has proven to be very helpful in scoping and discussing expectations throughout the different departments of the library [2].

3. PRESERVATION LEVELS

An important requirement for the New e-Depot is that it should be capable of processing and managing multiple digital collections at different preservation levels. Not every collection represents the same value to the library, not every collection is preserved for the same reasons and not every collection needs the same treatment to ensure permanent access. If all digital publications have to be processed and managed at the highest quality level, the digital archiving environment would become unaffordable. KB therefore defined a set of preservation levels and a value- and risk methodology to link preservation levels to digital

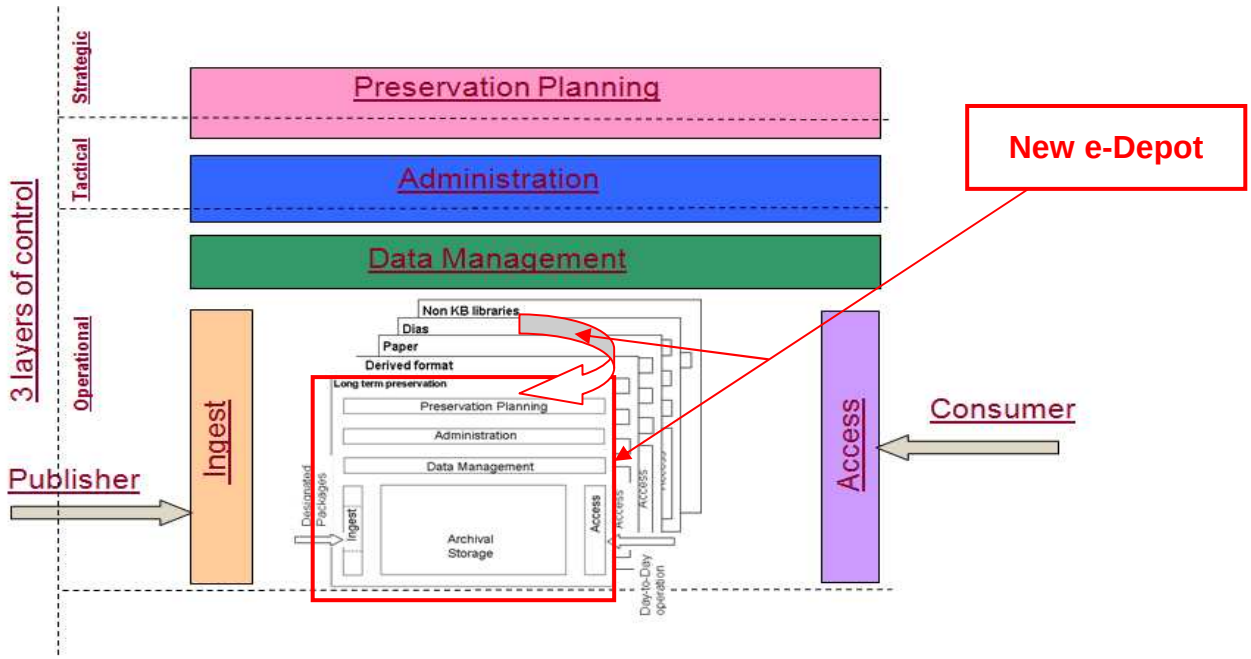


Figure 1. Two-layered OAIS model

collections. This policy has not yet reached its approved version, but the general outlines are clear and have to be put in practice by the New e-Depot system (see [2] for more details).

The preservation levels will consist of:

- Level 0 for collections that will not have to be preserved by the library and will only be stored in a presentation environment;
- Level 1 or 'limited' level for collections that have to be preserved for more than five years but will not need to be fully checked on ingest and do not need large scale investment on preservation actions;
- Level 2 for collections that do have to be preserved for the long term (is more than five years), need to be checked on ingest but do not require future access in original file format. These collections will be subject to validation, will be stored on preservation storage, but may require less preservation actions;
- Level 3 for collections that have to be preserved for more than five years, need full ingest validation and preservation actions that secure future access in an authentic way.

4. DEFINING THE COMPONENTS FOR THE NEW E-DEPOT

Based on eight years of experience running the current e-Depot system, on international discussions, on working on digital preservation research projects and on growing insight into the processes that need to be supported, the KB team defined a components-based architecture for its New e-Depot environment. Three basic considerations have led to this set-up.

First of all, digital preservation is not just a matter of identifying technical requirements for secure storage. Far more than that it is the holistic approach of an organisation to achieve its preservation goals. It is defined by the services that institutes deliver, by checks on the publications on ingest, by management of information on the objects and the processes, by closely monitoring ICT developments and assessing what these developments mean, by storage management and the overall architecture of the preservation environment. So it's not just a few extra things you do after you store digital objects, but it is inherent to the organisational approach, the work processes and the automated steps that process, store and use digital content.

This leads to the second consideration, where preservation functionality is to be seen as an addition to more general requirements for a processing and storage environment. As digital preservation is a result of organisational approaches, work processes and systems, not all preservation functionality depends on specific systems. A long-term storage environment has perhaps extra features but is also 'just' a storage environment. Ingest for a long-term preservation system does include extra functionality but is also 'just' a processing workflow. It can very well be that standard IT solutions can deliver most of the required functionality.

Thirdly, a components based set-up allows for more flexibility, avoids vendor lock-in and makes it possible to choose the best product for each part of the archiving environment. As the KB is experiencing at this moment replacing a complete and integrated digital archiving system at once is a very challenging task. Choosing a modular approach will allow the KB to extend and improve the new system one module at a time. In the future, components must be replaceable by modern technologies more easily. This will add to the stability of the systems and avoids changes to stored content and metadata. Which brings us back to the first consideration, digital preservation is more than secure storage alone.

Based on these considerations, the KB team started to 'break down' components of a processing and storage environment into separate processes and services and made a translation into what we started to call 'real world building blocks' (not to imply that a library is not the real world...). By defining a combination between generally available IT components and special requirements for digital preservation, it became possible to set up an approach that would allow us to make optimal use of (commercial or open source) off-the-shelf software together with preservation focused services.

Considering the specific characteristics of the work processes that need support from the New e-Depot and quality attributes such as performance, adaptability, resilience to interferences and stability, the building blocks for the New e-Depot were chosen as depicted in figure 2. Each of these building blocks or modules will be described in more detail hereafter.

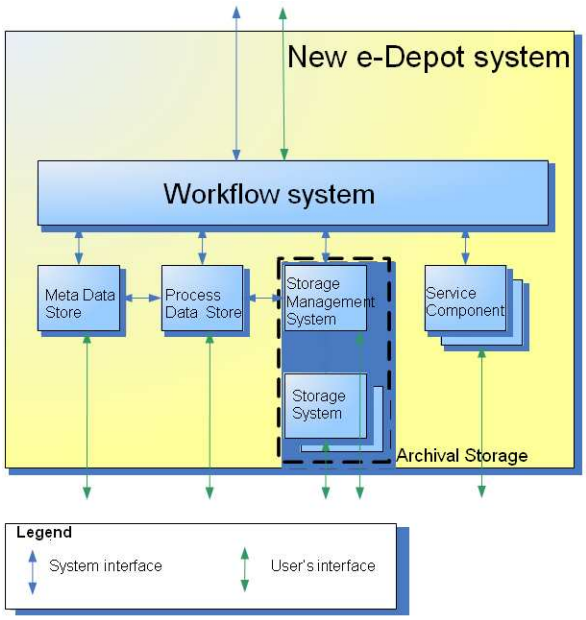


Figure 2. Building blocks New e-Depot

5. ARCHIVAL STORAGE MODULE

Archival Storage is provided by an implementation of a two-layered storage-solution. A Storage Management system abstracts other system components (more specifically the workflow system) from the actual storage provided by Storage Infrastructure (consisting of various storage media and network components).

The storage infrastructure will ensure that files are written and read to the actual storage media, in this way holding the enormous volumes of the actual bits of the material to be preserved. Because of the enormous and ever growing volumes and the crucial role in the archiving system, the storage infrastructure will have to have the following characteristics:

- highly reliable;
- scalable to very high volumes;
- cost effective;
- self monitoring.

Cost effectiveness plays a role for each of the modules, but for the storage infrastructure it is of the utmost importance. The volume of data will only increase year by year. The investments and operational costs for the storage media are by far the highest cost factor in the archiving system and will have to be controlled.

The Storage Management layer has an essential role within Archival Storage. It stores and retrieves files based on assigned unique identifiers, since access can not be based on storage locations or filenames. The reason is that the lifecycle of storage locations (e.g. the precise storage infrastructure and used media) and storage methods (defining constraints to filenames and locator structures) will undoubtedly change. These changes can be driven by the storage management system itself, when creating replica's of the original content on other storage infrastructures to safeguard the content from loss through hardware errors or disasters. Storage Management will also move large volumes of data to new storage infrastructures when older infrastructure or media are phased out. Storage Management should abstract all other system components from future technological changes. If this does not function properly, the stored data may well be perfectly retained in the storage infrastructure layer, but may be no longer accessible.

Another important requirement for a storage management layer is that it does not lay any restrictions on the storage infrastructure that is used in combination with the storage management layer. Therefore a storage management solution will have to be:

- highly reliable;
- independent of underlying storage infrastructure;
- implements a well described method and data store to connect content identifiers to storage locators.

6. WORKFLOW MANAGEMENT MODULE

The processes needed for ingest, access and preservation actions are provided by a Workflow Management system. This system implements ingest, access and preservation functions as defined workflows. It consists of a Process or Orchestration layer, a Mediation layer and a Transport layer to connect all systems to the workflows. The fourth layer is the service layer, that uses Service Components to implement specific atomic functionality to perform amongst others content analysis, content transformations and metadata conversions. The workflow system effectively implements the integration layers of a Service Oriented Architecture (SOA) [5]. This also implies that the workflow system will offer the entry point for all integrations with external systems.

Of the three types of processes the workflow module will have to support, the ingest process will put the highest demands on the system. Each day, tens of thousands of publications arriving at the KB in a large diversity of submission formats, containing several different file formats, will have to be validated and, if necessary, normalised to a more generic format. Depending on the preservation level, during long-term management, several preservation actions will be performed on the material. Next to integrating and orchestrating services, the workflow system will offer functionality to prioritize and parallelize workflows and service executing, perform load balancing to

optimize resource usage and offer message persistence and workflow resilience services.

The Workflow module must be capable of:

- processing high-volumes of data;
- support multiple workflows dependent on content types, producer and required preservation levels;
- offer support for manual intervention and repairs of invalid content and metadata;
- run different workflows in parallel maximizing throughput and balancing system load;
- allows for restart and recovery of failed workflow instances;
- halt and automatically resume workflows when services are temporarily unavailable;
- minimizing the development effort to implement new ingest streams (workflows);
- support the easy integration of specific preservation tooling.

7. META DATA STORE MODULE

Conformant with the OAIS model, metadata is stored in Archival Storage with the content. This makes the metadata subject to preservation together with the content. However, it also makes the metadata difficult to use by services that support the preservation action and access processes. To allow direct access to metadata needed to control these processes, it is not only stored as files, but also redundantly maintained in a Meta Data Store.

The metadata stored within the Meta Data Store serves as access mechanism to the data objects stored in the Storage Module. Its data model therefore structures the relationships between stored data objects and their metadata to enable the retrieval of selected Archival Information Packages (AIP). The data model offers placeholders to store identifiers the outside world can use to request stored content, when needed in a specific version or variant. The Meta Data Store also holds provenance data on actions performed on content and versions created.

The Meta Data Store will offer reporting functionality to query the systems database giving insight in the holdings of New e-Depot. To create these reports all stored metadata attributes can be used in queries and to structure the report. This report will be used as input in the preservation planning process, driving decisions on which preservation actions need to be performed.

The Meta Data Store will only hold a minimal amount of descriptive (or bibliographic) metadata and will therefore not be used directly by end users for requesting content stored in the New e-Depot. A separate bibliographic cataloguing system is available to search for content. Identifiers will be used to link the Meta Data Store of the New e-Depot with the external cataloguing system.

The relationships between stored metadata in the archival storage module and the data in the Meta Data Store are defined in such a way that the Meta Data Store can be rebuild when a disaster occurs using the metadata stored in the Archival Storage.

8. PROCESS DATA STORE MODULE

The purpose of the Process Data Store is to support the Monitoring & Control process for the New e-Depot system. More precisely:

1. it provides information on the execution of processes in terms of:
    a. measuring process execution results in a certain period to enable reporting on Key Performance Indicators (KPI's);
    b. reporting of deviations from the normal flow of operations (relative to defined benchmarks and tolerances);
2. it provides information to analyse trends in the growth and evolution of the collections processed and stored;
3. it provides information to merge process results with collection metadata, thus enabling analysis of the collections and related processes;
4. it provides information to sustain the integrity of all collections stored (both AIPs and Descriptive Information);
5. it supports consistency checking between the Meta Data Store and Archival Storage.

The Process Data Store does not provide the daily monitoring and control of the operational processes which are the responsibility of each system that performs or supports that operational process (i.e. primarily the Workflow Management Module).

The Process Data Store collects and receives data from other modules of the New e-Depot system and transforms and integrates them for reporting purposes. There is no automated feedback loop to these other modules and none of those other modules will depend on the Process Data Store for its proper functioning. The output of the Process Data Store will be used by operators and management for monitoring and control purposes.

9. DEVELOPMENT OF THE NEW E-DEPOT

Each module for the new system has been defined in detailed specification of requirements. On top of that, an overall architecture and data model have been designed. After a final review of the requirements, the procurement and development process will start in June 2010. A request for each component will be placed in the market separately and is expected to be filled in

differently. While workflow systems are widely available, both commercially and open-source, storage management is more specific and will see a different number of possible applications. Modules will either be bought, integrated or developed. The success of the approach will be largely defined by how the modules will be integrated, with each other, but also in the KB infrastructure. During the next few months, after the choice for applications and development of services has been made, it will be decided how the integration will be managed. In general, such an integration will look like as depicted in figure 3.

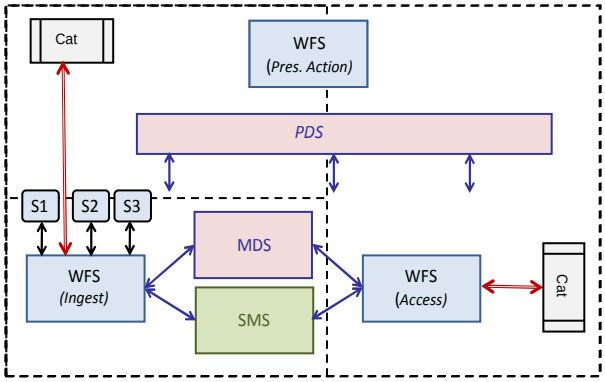


Figure 3. **New e-Depot integration**

The workflow system will be present in several processes (ingest, retention / preservation, access) and interacts with the New e-Depot modules Storage Management, Metadata Store and Process Data Store, and with external systems such as existing cataloguing services and possibly others.

10. REFERENCES

[1] KB Strategic plan 2010-2013, available at: http://www.kb.nl/bst/beleid/bp/2010/index-en.html (accessed 9 July 2010).

[2] Wijngaarden, H. v., *The seven year itch. Developing a next generation e-Depot at the KB.* Paper to be presented at the World Library and Information Congress, Gothenburg August 10-15 2010, available at: http://www.ifla.org/files/hq/papers/ifla76/157-wijngaarden-en.pdf (accessed 9 July 2010).

[3] IBM's Digital Information Archiving System (DIAS), available at: http://www-935.ibm.com/services/nl/dias/is/implementation_services.html (accessed 9 July 2010).

[4] *Long-term Preservation Services – a description of LTP services in a Digital Library environment*, Long-term Preservation Working Group, white paper will be available in July 2010 at: http://www.kb.nl/hrd/dd/index-en.html (accessed 9 July 2010).

[5] Bell, Michael (2008). "Introduction to Service-Oriented Modeling". *Service-Oriented Modeling: Service Analysis, Design, and Architecture*. Wiley & Sons. pp. 3. ISBN 978-0-470-14111-3

GUIDING A CAMPUS THROUGH THE TRANSITION TO A PAPERLESS RECORDS SYSTEM

Heather Briston **Karen Estlund**

University Libraries, University of Oregon
1501 Kincaid St.
Eugene, OR 97403-1299

ABSTRACT

The "paperless office" concept has been around for decades, and many have cited that the electronic office has instead increased the amount of paper produced. Case studies have shown that a successful "paperless" system requires motivation, ease of use, and cost savings [5]. Paper will co-exist with electronic records for the foreseeable future; however, what happens when the official record of an institution becomes "paperless"? This short paper presents a case study describing the efforts in the University of Oregon Office of the President to move to a fully electronic records system, the trickle-down effect to campus units, and the work of the Libraries to preserve the institutional record. The Libraries created a model to solve the immediate needs of the Office of the President addressing issues of workflow and preservation before an ideal system and staffing could be realized. A hands-on approach was employed, focusing on day-to-day work and ease of use for office contacts, and standards and migration plans for archival files using PLATTER [1]. By doing this, a foundation was created for an electronic records system that can be adapted across campus for administrative offices, faculty scholarship, cultural museums, science labs, and student coursework.

1. CAMPUS ENVIRONMENT

Records management at the University of Oregon (UO) has been mixed between a paper and electronic records system for many years. The University of Oregon has a long and proud history of decentralized information services and procedures, and does not require many specific systems be used across campus. As a public university, the institution's records must be kept in accordance with Oregon University System rules [2] and state public records laws [3]. Under the Oregon Administrative Rules that govern digitized and electronic records, born digital records can remain in their electronic form for preservation of electronic copies. For digitized records of permanent value, current rules require preservation in paper or microfilm. These rules are currently under review by the state. The University Archives, located within the Libraries, administers the permanent records of the University.

Beginning in 2006, many of the campus administrative offices such as Admissions, Registrar, and Financial Aid began using an enterprise document imaging system, Singularity, which interfaced with the campus-wide data management system, Banner. Yet, while all of this was occurring there were also homegrown and stand alone document imaging projects and data management systems being created throughout the university. From the perspective of electronic records management, while the document imaging system incorporated records scheduling into its infrastructure, most of the other systems existed with no plan or system for destruction or preservation. In most cases there was a reluctance to tackle this issue within departments because of the enormous scale and the dearth of available resources. Prior to the 2009 effort, except for occasional final reports received in digital form and made available through the university's institutional repository[1], there was no plan or workflow for comprehensively collecting and preserving the electronic records produced by an office. In only one prior instance was this done: a small office in International Affairs closed and its records concerning the events surrounding the granting of an honorary degree were transferred to University Archives on floppy disks. Lacking a workflow or storage space the files were converted to PDF and put in the institutional repository. The native files were put in a dark archives. Subsequently there have been challenges with providing context and identifying the files as archival, rather than current in this online environment.

At the University of Oregon, the President is the chief executive officer of the university. During this period, there was widespread use of Microsoft Office products within the Office of the President, including Outlook for e-mail and calendaring, but for preservation purposes all important records were printed in triplicate and filed in chronological, topical and high profile issue files. There was no integration of a digitization project for paper or

[1] Scholars' Bank, http://scholarsbank.uoregon.edu

preservation efforts for the born-digital electronic records within the office until after the close of the presidency.

2. CHANGE

With the arrival of the new university president on July 1, 2009, there was a new focus on electronic records produced by the Office of the President. Only one of the previous executive assistants remained with the new administration. New Office members and administrators possessed a greater facility with the use of technology in records creation. Efficiency and use of technology to improve efficiency was emphasized. As a result, not only were important documents not printed in triplicate, but the Office, under the direction of the President, committed itself to going paperless, scanning any documents received in paper format and refraining from printing except when required.

In the previous environment, records were delivered to the University Archives annually. The transfer was routine and institutionalized. With the change to electronic records, there was opportunity both to lose access to records that formerly would have been delivered in print, but also to gain access to records such as email that may have been left out of the transfer to the Archives, as well as receive files with the original file metadata appended.

Prior to his arrival at UO, Richard Lariviere was the Provost at the University of Kansas (Lawrence). While Provost he oversaw the start of a campus wide, comprehensive information management program, which brought together digital information security, electronic records management and archives, as well as digital asset management and preservation.[2]

Conversation on campus has quickly disseminated the Office of the President's new emphasis on reducing the reliance on paper and improving efficiency in university administration. Other offices on campus have actively contacted the University Archivist to seek direction on how to better manage and access their records electronically and ultimately to schedule and transfer electronic records for preservation. University Development is creating a library of documents for access throughout their offices, encouraging people not to print additional copies. The College of Education is implementing a system to manage the creation and management of grants and other financial records in electronic form. The College of Arts and Sciences is hiring a records assistant to help them manage their records in all formats.

3. OFFICE PROCEDURES

To fully launch an electronic records system in the Office of the President, the Office personnel have begun investigating records management systems to meet their needs and legal requirements. In the meantime, records are still being created, and the Office is not waiting for the perfect system before transitioning. To help ease this transition into an electronic records system, ensure that standards are met, and that files may be easily transferred to the chosen system with preservation in mind, the University Archivist and Digital Collections Coordinator met with Office staff in fall 2009. This provided an opportunity to jointly conceptualize a campus workflow for the transition of electronic records to University Archives and to identify unforeseen problems.

The goal of the initial meeting was to advise on procedures for turning working documents into records. Three main topics were explored:

1) Migration of working files to records (including reformatting)
2) File naming conventions
3) Tags and categories to easily retrieve relevant documents

3.1. File Migration

The staff of the Office of the President, unsure of how to proceed, had hybridized practices from the last administration and the goal of going paperless. They were printing out electronic documents, rescanning them into PDF files and then storing them on the networked file shares. By using tools already at their disposal and creating brief instructions, the Archivist and the Digital Coordinator were able to demonstrate how to create full text searchable PDF files from Word and other documents using Adobe Acrobat Pro. Because the Office is not completely paperless, staff were instructed in Adobe Acrobat Pro's native Optical Character Recognition (OCR) engine so that scanned documents could also be made full text searchable.

Primarily records created by the Office are Microsoft Office documents, but there are also digital audio / visual files, digital photographs, and web based records. In order to maintain functionality in Microsoft Office documents, such as the searching and tags in Microsoft Outlook, staff were taught to transfer native file formats to PDF. If there was concern that particularly sensitive information might be changed or modified, a PDF was requested for submission to the Archives as well as the native file format. This allowed the staff to feel more comfortable with the transfer of editable file types.

3.2. File naming

The staff in the Office had a good initial sense about how to uniquely identify files so that they could be easily retrieved. Their work environment demands that they be able to quickly retrieve items as needed; therefore the general principles of uniqueness and easily recognizable file names were already in place. The file names, however, had many special characters and spaces. The staff easily understood that these might cause problems. Staff were

[2] http://www.provost.ku.edu/infomanagement/index.shtml

introduced to a simple Freeware tool, ReNamer,[3] with the ability to mass apply file naming changes and strip out unwanted characters. A limit of 15 characters to file names was suggested whenever possible.

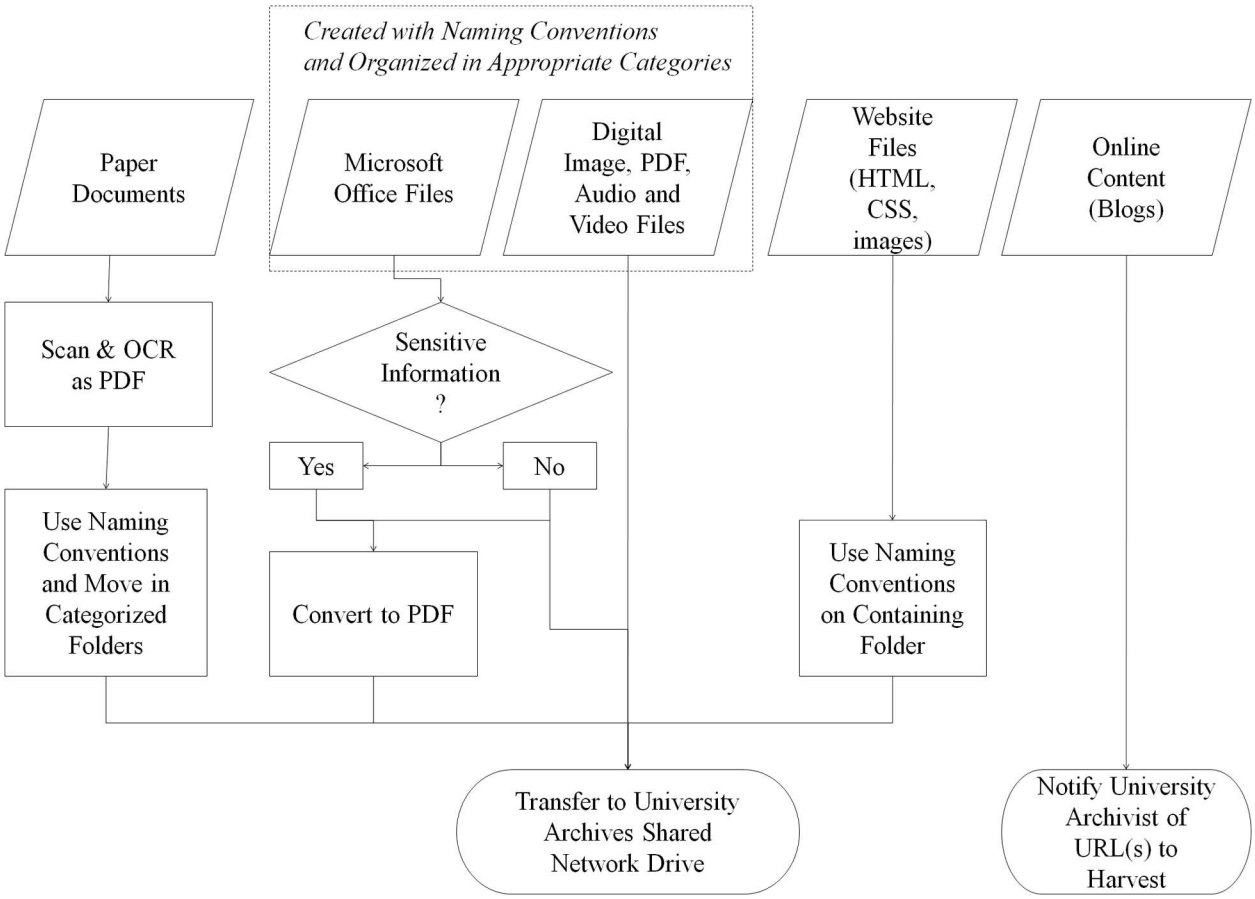


Figure 1. Office of the President transfer procedures to University Archives.

3.3. Categorization of Files

The most exciting part of electronic records for the Office staff was the ability to tag and categorize files without having to make triplicate print copies. This was especially valuable in the area of email, where utilizing the tags and flags in Microsoft Outlook could help easily retrieve relevant emails. The staff have begun to make lists of their desired categories in consultation with the University Archivist and Digital Collections Coordinator. The goal is to create a standard list of category names. Examples of these categories include:

- Correspondence
- Reports
- Speeches
- Athletics
- College of Arts and Sciences, etc.

4. PLANNING FOR PRESERVATION

Ideally, any future repository system will be based on the Open Archival Information Standard (OAIS) [4] or at the very least the campus will use a single records management system. In the case of the Office of the President, they had yet to identify a records management system to use. In order to facilitate a workflow that could quickly be constructed for secure ingest, the principles of the OAIS model were followed as closely as possible, with manual controls in a simple file system infrastructure linked to descriptive records in Archivists' Toolkit.[4] The infrastructure was constructed to easily allow migration to an OAIS compliant repository in the future.

PLATTER documentation was used for strategic planning [1] to help express goals and plan object ingestion, migration schedules, institutional support,

[3] http://www.den4b.com/downloads.php?project=ReNamer

[4] http://www.archiviststoolkit.org/

technical infrastructure, and access conditions. With theory and planning in place, various elements were implemented on the path towards a preservation system. These administrative steps were essential in defining the roles of the local office and those of the Libraries, as well as the infrastructure needed.

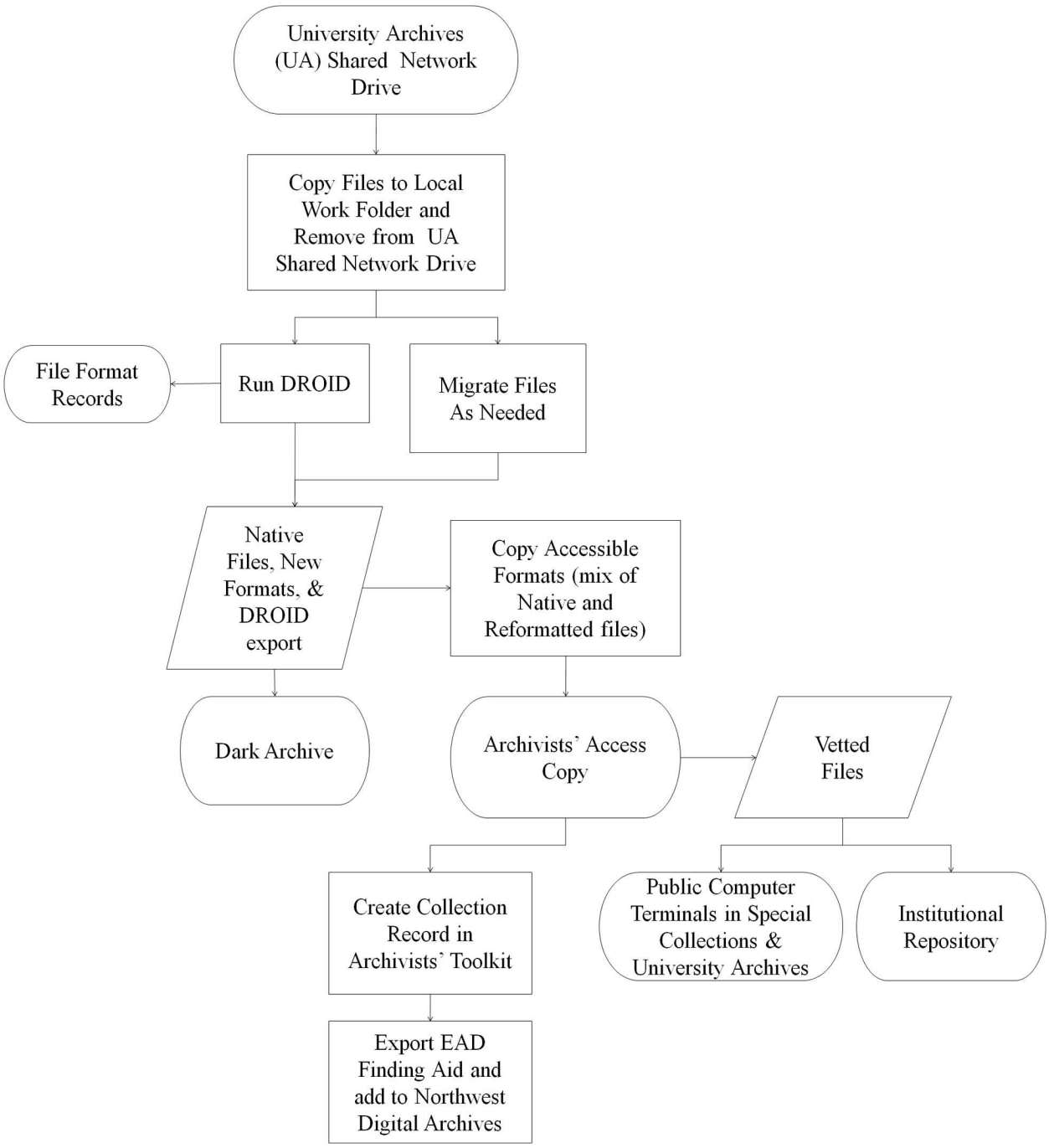


Figure 2. University Archives processes for preservation and access of electronic records.

By using familiar tools available to the staff of the Office of the President at the point of document creation, advising on protocols for file naming and description, creating easy ingest mechanisms through mapped network shared drives, an easy-to-implement workflow from a campus office to University Archives

was defined. This temporary storage is divided into folders for each campus department and access is restricted to staff in those departments, the University Archivist, and Library IT.

Once in University Archives, digital preservation strategies put in place by the Libraries were integrated and access provided through tools currently used by Libraries and Archives staff. The files are inventoried using DROID (Digital Record Object Identification)[5] and file formats are converted if needed. (For example: video files are converted to .mp4.) A text file is exported from DROID and saved alongside the native and converted files and transferred to a dark archive with bit level integrity checking and backed-up onto magnetic tape. The lists generated by DROID are also kept in a central location, which acts as a store for all file format lists and is monitored for assessing any necessary future file migrations. A second copy of the file is available as the Archivists' access copy, which can be modified and re-categorized, with the collection record in Archivists' Toolkit pointing to this location. An EAD (Encoded Archival Description) finding aid is exported from the Archivists' Toolkit and added to the Northwest Digital Archives.[6] Publicly accessible records vetted by the University Archivist and the UO Public Records Officer will be made available on computer workstations in Special Collections and University Archives and/or uploaded to the university's institutional repository.

The solution is far from perfect and could not pass a trusted repository audit; however, it is a first step in the implementation of an electronic records program and the beginnings of a comprehensive plan to preserve the full history of the current institution electronically.

5. THE CULTURE OF CHANGE

The growth in acceptance of managing electronic records and the validity of the electronic record as a "record," has quickly spread across the UO campus. Since the arrival of the new president, new efforts are materializing to use technology and new electronic systems and to preserve the output for the future. The motivation is not on using the technology alone but on what advantage the technologies provide. This emphasis will be key to the implementation of a fully paperless records system, and it is the responsibility of the University Archives and the Libraries to ensure that it can be preserved.

5.1. University Senate

In order to involve members of the campus more widely in campus governance, at the initiation of the current University Senate president and executive committee, the final three Senate meetings of the academic year will be captured in digital video and streamed for wider viewing. As the minutes and other documents capturing the activities and decisions of the senate are considered permanent records, the recorded senate meetings will be retained and preserved by the University Archives.

5.2. Teaching and Students

There is increasing use of Web 2.0 tools for collaborative student learning on campus; most of it is ad hoc, driven by faculty and pedagogy, or in some rare cases, student influence. One result is the creation of blogs for e-portfolios, particularly in business classes and architecture, and the potential for campus wide multi-user blogs for students and faculty. These campus departments are seeking advice from the Libraries on how to preserve these records.

5.3. Faculty Scholarship

The University of Oregon created an institutional repository (Scholars' Bank) for faculty scholarship in 2003. Like many institutional repositories it has had continued but limited use by faculty. This year the Department of Romance Languages mandated that their faculty deposit electronic versions of their scholarship in Scholars' Bank.

The campus science faculty have also begun to think seriously about the preservation of data they create. Although most do not wish to contribute their data directly to the University Archives, they are seeking guidance on preservation issues, formats, and especially metadata and description from the Libraries.

5.4. Museums

The Jordan Schnitzer Museum of Art and the Museum of Natural and Cultural History at the University of Oregon have begun looking beyond online exhibits and using digital images only for their own internal searching. Previously, the museums retained all their data and digital images on hard drives next to a work station in the building. The Museum of Natural and Cultural History has been in conversations with the Libraries on proper image formats, file naming, and back-up and storage for their data. They recently hired a "Conservator and Digital Archivist" to help in this process.

6. CONCLUSION

With the motivation provided by the new university president, the Libraries is able to assist in an easy to use system and on the way to making a paperless records system successful. The University Archives and Libraries are quickly adapting methodologies, standards, and procedures to ensure the preservation of these

[5] http://sourceforge.net/projects/droid/ from the National Archives of the United Kingdom, PRONOM.

[6] Northwest Digital Archives (NWDA): http://nwda.wsulibs.wsu.edu/

materials. We cannot wait for the perfect system or uniform systems to be used across campus. By adapting the conceptual standards of digital preservation and an easy-to-adopt workflow, we will be able to guide the campus through the change to electronic records.

7. REFERENCES

[1] DigitalPreservationEurope, (April 2008), "DPE Repository Planning Checklist and Guidance DPED3.2", http://www.digitalpreservationeurope.eu/publications/reports/Repository_Planning_Checklist_and_Guidance.pdf

[2] Oregon Administrative Rule, Secretary of State, Archives Division, Oregon University System Records, http://arcweb.sos.state.or.us/rules/OARS_100/OAR_166/166_475.html

[3] Oregon Revised Statutes, Chapter 192, Records, Public Reports and Meetings (Public Records Law), http://www.leg.state.or.us/ors/192.html

[4] Reference Model for an Open Archival Information System (OAIS) CCSDS 650.0-B-1 Blue Book (January 2002), http://public.ccsds.org/publications/archive/650x0b1.pdf

[5] Sellen, A. J. and R. Harper. *The Myth of the Paperless Office.* MIT Press, Cambridge, Mass., 2002.

Session 9b: Case Studies

FULL PAPER

REPRESENTATION OF DIGITAL MATERIAL PRESERVED IN A LIBRARY CONTEXT

Eld Zierau
The Royal Library of Denmark
Dep. of Digital Preservation
P.O.BOX 2149
1016 Copenhagen K, Denmark

ABSTRACT

This article explores preservation of digital material in a library context with a focus on logical object modelling that takes both preservation and dissemination into account. The article describes normalisation of data expressed via a logical object model. This logical object model is designed to support the requirements for joint preservation and dissemination. Additionally the article includes a suggestion for a possible implementation that respects the logical object model.

Formulation, of the requirements and possible implementation for a logical object model, is based on observation of current trends, as well as results from a research project on preservation strategies for libraries. The research project has been carried out at the Royal Library of Denmark, and it is based on a case study of a 10 year old web application containing the Archive of Danish Literature. The formulated requirements include e.g. requirements for many-to-many migration in preservation and requirements for homogenous navigation and social networking in dissemination.

Many of the described observations and results have parallels to other types of material. These parallels are partly described, and thus the results can be used as a contribution to development of systems and strategies for preservation and dissemination in the new decade and beyond.

1. INTRODUCTION

This article explores digital preservation in a university and national library context where preservation must go hand in hand with dissemination. It focuses on the object modelling aspects to represent a normalisation form that supports future functional preservation as well as dissemination. Functional (logical) preservation here means preservation of a digital object to ensure that it remain understandable and usable on a long term basis. The study is a result of a research project at the Royal Library of Denmark (KB). The goal is to investigate preservation strategies in a library context.

The hypothesis investigated is that it is possible to reuse and normalise existing data from digitisations (10 years or older). If this is the case, it will be economically beneficial to preserve the normalised data in the sense of preserving the investment of the earlier digitisations. The results of exploring the hypothesis will influence the future normalisation of data as well as preservation and dissemination strategies.

The research is based on a case study of the Archive of Danish Literature (ADL) system. ADL is a web-based framework constructed at the start of the century. ADL is mostly limited to books, book collections and book metadata, but parallels to other types of material can be drawn. A separate part of the research project investigated whether the original digitised ADL was worthy of preservation for future use (study part 1) [5], which the study found to be the case. The other part of the study is the one presented here. This part will only look at the normalisation and logical object modelling aspects for the digital material and their data structures.

In our view preservation and dissemination are highly interrelated. This leads us to assume that they must be managed jointly on a day-to-day basis regarding ingest, access and maintenance, as illustrated in the Figure 1. The terms used here are defined in the OAIS reference model[1], unless an explicit definition is given.

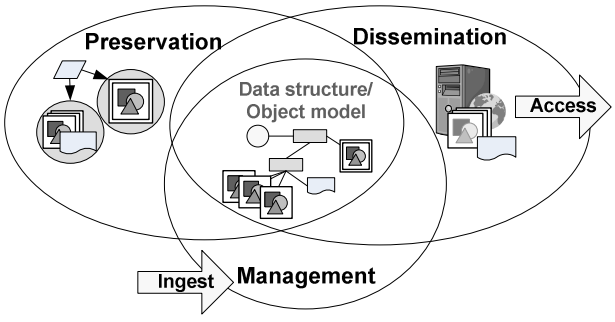


Figure 1. Preservation and dissemination interrelations.

The background for this view is that libraries have an obligation both to preserve and disseminate material. This fact challenges the demands on preservation, where

[1] OAIS (Open Archival Information System). 2002. ISO 14721:2003.

material in many cases must retain a short and efficient route to dissemination through fast access by the public or researchers and in a user friendly way. Both dissemination and preservation demands are under constant challenge as a result of technological evolution. New requirements emerge such as representations to new media e.g. mobile devices, representations in new form e.g. e-books[2] or high resolution images, and representation information via social network communities[3] [1]. This means that digital material becomes more inhomogeneous with new representations. Furthermore, the need for different preservation levels becomes more apparent. Ten years ago, the focus was primarily on digitised books, while we today face challenges with e.g. contents from a PC of a deceased author, internet harvests, emails, and digitised images from deteriorating negatives [2].

The purposes and goals for dissemination and preservation are different. Their interrelation means that the requirements for dissemination need to be taken into account when we formulate the long term preservation strategies. Furthermore there are requirements to allow for data migration into preservation formats with different storage characteristics. Migration will here mean modification of the digital objects to ensure permanent access to these objects. The storage characteristics can be: how much storage space the format requires, or how different parts of a logical object e.g. a page image, are stored with different confidentiality levels and different bit preservation levels [7], i.e. different bit safety levels ensuring that the actual bits remain intact and accessible at all times. Most of these requirements must be taken into account when we define an object model for normalised data.

Before we can describe an object model for normalised data, we will list the relevant dissemination and preservation requirements based on the case study, the experiences gained, and the relevant results from study part 1. Some of the requirements will relate to an actual system implementation. This article will therefore include a description of a possible solution for digital object management systems (DOMS) that can support workflows of ingest, ensuring preservation and dissemination of the digital material of a library. The possible solution description is based on results from a DOMS pre-study at KB carried out by joint forces from the Digital Preservation Department and the Digital Infrastructure and Services department at KB.

2. CASE STUDY: THE ADL SYSTEM

The ADL System is used as a case study, in order to study new requirements for dissemination and preservation that emerged as a consequence of the technical evolution in the last decade. The case study is interesting because it reflects a system built on the basis of technologies from the start of this century. The case study gives us indications of the challenges to take into account when we consider a future DOMS, regarding present requirements, and regarding trends that should be addressed for future requirements. Although the ADL system is a case study covering specific materials, the indications will have parallels to other types of material. When the requirements are specified in the next section, such generalisation will be made where possible.

2.1. Short Description of ADL

The ADL system was developed by KB together with "Det Danske Sprog- og Litteraturselskab" (DSL) which publishes and documents Danish language and literature. KB developed the framework, while DSL selected literary works to be included. The system is a web based dissemination platform for digitised material from the Archive for Danish Literature. Today it contains literature from 78 authors represented represented by over 10,000 works of literature (defined as a work by an author that can represent itself without other context, examples are novels, poems, plays). ADL additionally contains author portraits as well as 33 pieces of music (sheet music) and 118 manuscripts. The publication framework is still available on http://www.adl.dk/.

The structure and design of the underlying ADL database is based on book pages, authors, their literary works and the period when the authors were active.

Since ADL was designed a decade ago, its navigation and search facilities along with design of data structures are old-fashioned compared to the possibilities of present technology. Although ADL has served as a good application, it now needs renewal which will partly be specified on basis of the research results.

2.2. Experiences from ADL

The ADL system does presently offer separate views of book pages in three ways based on three different digital representations of the pages, but there are no relations between the views. The views are: a 4-bit GIF image, a pure text representation, or a page can be downloaded as a PDF file containing the page image for print.

The data structure is highly dependent on pages, which gives several challenges. The structure of page images in a book is specified in a TEI-P4[4] LITE XML. The XML is uploaded to a database which is used for dynamic generation of HTML pages. The page number is used in the name of the related page files with page image and encoded text. This eases application coding of references to different representations of a page in GIF, text or PDF, but introduces a number of challenges. Firstly it challenges maintenance if page numbering needs to be corrected, not only should the file name be changed, but all references from e.g. citations via hardcoded URLs will need update as well.

[2] See http://en.wikipedia.org/wiki/E-book

[3] As define on http://en.wikipedia.org/wiki/Social_network_service

[4] TEI (Text Encoding Initiative).

Another related challenge is that there can exist different versions of a page image. For example, ADL had a copyright restriction on illustrations appearing as part of a page. This restriction was only enforced within a certain period, thus two versions exists for such pages, both in the GIF image, and in the PDF derived from the original TIFF image. File names with page numbers will also cause problems for functional preservation that are similar to the problems of preserving web archives[5].

The navigation and search facilities depend on older technologies and the data structures. Limits became apparent in particular for navigation, when sheet music in PDFs with JPEG images were added (originally digitised for another purpose), and when manuscripts represented in JPEG files (for better dissemination of colours) were added. One problem was that this new material is not viewed as literary works and therefore did not fit in the original navigation structure. Furthermore navigation of sheet music between pages is different, since they are fully represented in a PDF file.

Inclusion of additional material information in ADL has made the lack of referencing possibilities apparent. Examples are reference to other resources on external web-sites, or Danish translations for books written in Latin. The original ADL data model was not designed for these inclusions and they are therefore not logically integrated in ADL, e.g. the translations are hard to find and the relation to the book is not obvious.

A rare challenge in ADL occurred when a literary work, in the form of a novel, was added. The challenge was that the novel was represented in two volumes. The solution was to represent the two volumes as one book in ADL, with one XML file for both volumes.

ADL has an option for users to send an error report on errors in the OCR text. A challenge here has been to have dedicated time to handle the error reports, which are handled manually. Furthermore, the current ADL system does not have automatic version control on changed text, thus the changes can be hard to track.

Presently, the ADL is only preserved as a part of the Danish web archive. That means that the only data preserved is the data visible on the internet, which does not include e.g. special encoding of texts. Further actions for preservation await the research results.

2.3. Relevant Results from Experiments

In connection with the study part 1, we have done experiments involving two re-digitisations. Some of the results from these experiments also influence the normalisation considerations, therefore we here provide a short summary.

The re-digitisation was carried out in two places and with two different approaches. One carried out a mass digitisation including new scanning of the books (referred as SC1). Another used an approach similar to the original ADL digitisation (referred as SC2).

A conclusion from study part 1 was that the original ADL scans were worthy of preservation. There were, however, cases of missing pages in the ADL scans. The missing pages were mostly blank pages or pages with editorial information, but in one particular case, the missing pages contained parts of a poem. This gives an example of a case where we would like to add page images from the new SC1 scan to the existing ADL.

Another conclusion was that the original ADL XML encoding was worthy of preservation, but additional results from SC1 and SC2 should be added and preserved as well. The additional results were the encodings for the missing pages, and the marginal notes which originally were left out in the ADL encodings.

Updating the encodings challenges the representation. One challenge is that the encoding results differed due to the different encoding formats. The differences are both in coverage and in type of XML tree structure. The ADL and SC2 XML are given in TEI-P4 per book and the SC1 XML is given per page in ALTO[6]. Positions in ALTO from SC1 refer to SC1 scans, while it is the ADL scans that are preserved. Thus if positions are added for future referencing mechanisms or creation of searchable PDF, we will need to produce this information based on the ADL scans. Lastly, the encoding of marginal notes is interesting, because the SC1 XML marking notes via positions was the most precise result. In the SC2 XML notes were marked notes with reference to a full paragraph, which is not precise.

3. REQUIREMENTS

On the basis of our knowledge of growing demands, experiences and experimental results, we can now describe the requirements for dissemination and preservation. These requirements can be applied for book collections in general, and for other materials.

3.1. Requirements for Dissemination

The technological evolution of the last decade has opened many new dissemination possibilities. For example, faster internet connections have made it possible and more common to have videos and high resolution images as part of web material. Digital born material like e-books is becoming more common. More advanced presentation in websites is appearing, e.g. synchronised representations with annotations possibilities[7]. Consequently, the requirements for the ADL application are increasing in accord with these new possibilities. The information we want to disseminate has evolved as illustrated in Table 1.

[5] See e.g. "Long-term Preservation of Web Archives – Experimenting with Emulation mad Migration Methodologies" on http://netpreserve.org/publications/ NLA_2009_IIPC_Report.pdf

[6] ALTO (Analyzed Layout and Text Object). 2004. Technical Metadata for Optical Character Recognition, version 1.2.

[7] See f.ex. http://openvault.wgbh.org/catalog/org.wgbh.mla:7376e451372c8a219648fc8e424aa9a1e8b463a4

Present ADL dissemination	**Extra desired dissemination**
Book page images (GIF-images, text, PDF download)	Other book manifestation of book item
Author citation	Content segments
Author description (picture, period, important dates), Period description	Thematic ontology Timeline with literary works
Sheet music & manuscripts	Other related material
Overviews (list of literary works, author list, period)	Time line, thematic ontology, student material, etc.
Error reporting option	Social network community (OCR correction, annotation, quiz etc.)

Table 1. Present and future dissemination.

The contents of Table 1 is based on generalisation of the current contents, on current technologies as mentioned above, and on new user requirements like plays in other manifestations, and social networking.

Generalisation of a book item is a book manifestation (item and manifestation concepts as defined in IFLA [3]). That means a manifestation in form of another edition, a translation, synthetic reading of encoded text, a live-recording of a play of a book containing drama, or it could be a manifestation in other dissemination formats like an e-book, a format for mobile devices etc.

Generalisation of citations is content segments, which can be an arbitrary part of the book, for example a chapter interval, a citation, a page interval, a literary work or the whole book. It must also support references that mark translated text, or references in connection with annotations, e.g. created by the public.

Other related author material can be anything from supplementary material to references to other dissemination platforms. Such material may also need to refer to parts of the material. For instance the sheet music may refer to a certain part of a play.

Social networking requirements are the most comprehensive generalisation of requirements. They are interesting for libraries, as a means to obtain corrections of digitisation, to get additional information on material, and to evolve interest groups as part of library life, for instance quizzes or student material related to the material [1]. Annotation may also come from research communities. An example is KB's involvement in the CLARIN project[8] which concerns infrastructure for scientific data. In CLARIN the ADL books are to be 'part of speech' encoded, where all words will be encoded with classifications of verbs, substantives etc.

[8] Common Language Resources and Technology (CLARIN). http://www.clarin.eu/

General requirements will still apply, such as scalability, fast response time, user friendly interface. These requirements deserve special attention for a future context, since the magnitude and variation of data collections are increasing, which challenge scalability and fast response time. User interfaces should be homogeneous when they cover similar material digitised and represented in different ways. Search facilities set requirements for indexing and search in collections that may cover a range of material from many existing web applications.

An additional requirement comes from the growing demands for simultaneous display of different views and their interrelation. An example is synchronisation between audio and text e.g. using DAISY[9].

3.2. Requirements for Preservation

Our requirements for preservation are based on a decision to preserve digital born material and digitisation material to be reused in a future context. The preserved material will be the basis for a transformation into emerging dissemination and preservation formats. The assumption of reuse is the reason why we here only will consider a migration strategy. Emulation[10] does not support changes in presentation form and is therefore not considered.

From a preservation point of view, normalisation should be as simple as possible, and based as much as possible on standards in order to ease future understanding. Many different standards can support a final implementation. Examples are PREMIS[11] which provides a standard for preservation metadata, METS[12] which provides a standard to express object structure. Implicitly this also means that preserved data must not be structured in order to suit specific tools.

We need a flexible data structure for functional preservation in order to be able to represent a book object and its different migrations in the form of digital objects including structural and technical metadata. Furthermore, the relation between representations can become complex in the future, since we already know of cases where there are many-to-many relations between the digital objects, for example, many digital page images versus an e-book. A requirement is therefore to have a flexible object model where such representations and many-to-many relations can be modelled.

Another part of functional preservation is to preserve references into material, like citations references or future annotations. The modelling must therefore take into account how references into objects can be migrated as part of a full migration. The modelling must

[9] See http://en.wikipedia.org/wiki/DAISY_Digital_Talking_Book

[10] See e.g. "Keeping Emulation Environments Portable" (KEEP). http://www.keep-project.eu/

[11] PREMIS (Preservation Metadata Implementation Strategies). 2008. Data Dictionary for Preservation Metadata, version 2.0.

[12] METS (Metadata Encoding and Transmission Standard). 2009. Version 1.8.

also allow creation of new versions with added contents as in the example of the missing pages.

Finally, it must be possible to store the data at differentiated confidentiality and bit safety levels, e.g. illustrations with copyrights have higher confidentiality than the rest, and the digital born material, such as author descriptions, needs a higher level of bit safety than the digitised book images, as long as the physical book is still available.

3.3. Interrelated Requirements

The interrelated requirements are the requirements derived from the interrelations between preservation and dissemination.

We will here view a logical object as a representation of an AIP (Archival Information Package) defined in the OAIS reference model. In OAIS, all preservation information is available in an AIP. However, not all information in an AIP is needed for dissemination. In OAIS the information for dissemination can be derived from enriched and transformed data.

We will require that logical object representations of the preserved data are relatively similar to representations in dissemination, and visa versa. The reason is that we will need to minimise processing time and storage cost for dissemination and preservation.

When we focus on storage, we also need to analyse possibilities for reuse of stored data between dissemination platform and the preservation platform. For example, if they both use the same high consumption storage formats, they can share one copy used as part of the bit preservation. Sharing a copy should however be done with care [7].

Another possible cost-reducing architecture could be that dissemination relies on cache storage with a possibility to retrieve preserved data on request. In this case preserved data must be easy to identify and retrieve. However, also in this case the transformation from a preservation representation to a dissemination representation must be minimal in order to meet time and scalability requirements.

Note that these last requirements can mean an indirect requirement of coordinated shift in the preservation and dissemination formats. An example could be that dissemination of book pages was changed from TIFF to JPEG2000, and similarly for preservation.

4. DRAWING LINES TO THE FUTURE

In this section we will suggest a flexible object model for normalisation of data objects and their metadata, which can meet our requirements for functional preservation in a library context. Additionally we will point at possible implementations in a DOMS, on basis of current state of the art of library DOMS', and architectural and community requirements.

4.1. Suggested Shared Logical Object Model

This section will present a flexible object model which enables us to normalise the data in a way that respects our requirements for preservation and dissemination.

The suggested logical object model is meant as an abstract model which is respected in the explicit implementations. That means representations for dissemination do not need to be implemented in the same way as representations for preservation, although they do need to meet the requirement to retain a short route to dissemination.

The logical object model is inspired by an initial object model from the Planets project[13] and the additional work with a concrete implementation including simple ER-diagram developed in the Pindar project [6]. These object models support functional preservation including many-to-many migrations.

4.1.1. Representations

The logical object model operates with different object representations. A representation must be a self-contained representation of the object, independent of other representations. Examples are representations of different migrations, different versions, different derived versions etc. This is exemplified in Figure 2. The example given in Figure 2 could be a future version of ADL material, where page images have been migrated to JPEG2000, but the corresponding dissemination format is JPEG. Note that not all representations are preserved, e.g. the JPEG. Other examples of representations that could be added are synthetic voice or an e-book version.

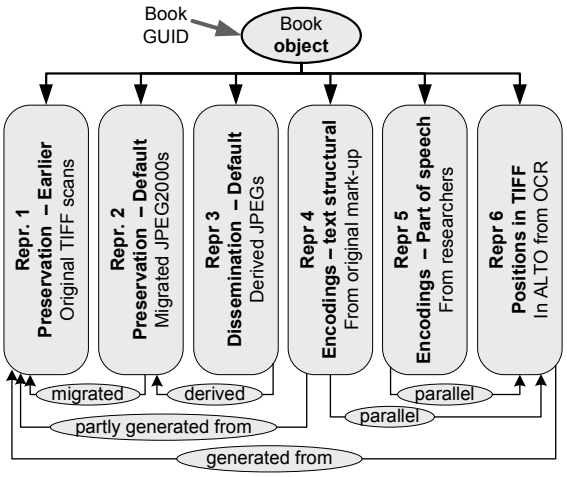


Figure 2. Example of representations of an object.

The different representations relate to each other in different ways. For example Repr. 6 was generated from Repr. 1 as part of the digitisation process. This is also the case for Repr. 5, but only partly, since it was enriched with manual encodings as well. For

[13] Preservation and Long-term Access through NETworked Services (Planets). See http://www.planets-project.eu/

preservation and reproduction purposes, the technical details on how e.g. a representation is derived must be part of the metadata in the same way as technical metadata for a preservation migration, as e.g. described in the PREMIS standard.

It is not part of the model to define what kind of representations that can and must be included. It only prescribes that their relations must be described in detail. This creates a possibility for addition of new representations. It also creates a possibility to have more migration representations for one migration, which will be the case if different aspects of the original file will need to be represented in two different formats.

The concept of having representation also makes it possible to define groups of logical objects with common behaviours, both with regard to preservation aspects such as migration, and dissemination behaviours such as presentation in e.g. a web interface.

In the example, there are different encoded text representations. This illustrates a choice of keeping a split between different encoded texts for preservation, e.g. for positions, part of speech and text structural encodings like chapters and stage directions in drama. This is especially preferable in a preservation perspective since the encodings are based on different parts of characters in the text, which will require encoding of overlapping hierarchies. This is a complex task, which contradicts the desire for simplicity in preserved data. Deriving and migrating information will therefore be harder, and there will be a risk of introducing errors in updates. Furthermore, positions may deserve separate representation, since they only make sense for a very specific page image, e.g. separate position sets may come over time, and some may loose value due to deletion of related pages. On the other hand a disadvantage is that the OCR-text may have to be in all encoding representations. Note also that, even though some complexity can be eliminated by splitting up the encoding, there will be aspects where we cannot avoid some overlapping structure, as exemplified in [4].

In a future dissemination perspective where we want a dynamic environment, with frequent changes in the encoded text as a result of social networking, it will be better to have one source of update, i.e. a representation with all encodings including all overlapping trees, e.g. in an XML database. Such a representation could be added, as long as thorough description of relations to separated encoding representations is described.

The fact that dissemination is extended to include ingest operations in the form of quality checked corrective and extension information via social networking, complicates the interrelation between dissemination and preservation. Most preservation actions, e.g. bit preservation, can only be done on static material, thus the dynamic aspects will need to be represented in snapshots. The ingest process part must therefore be carefully considered, especially, if the encodings are represented differently. Furthermore, there will be a challenge in having asynchronous representations where the dissemination representation may be more correct than the preservation representation, as a consequence of social networking information that has not yet been quality checked and ingested.

4.1.2. Detailed Logical Object Model

A detailed logical object model must respect requirements for representation of many-to-many relations, referencing into objects, and a possibility to make corrections, e.g. by adding extra pages. Figure 3 illustrates the detailed logical object model by some book representation examples.

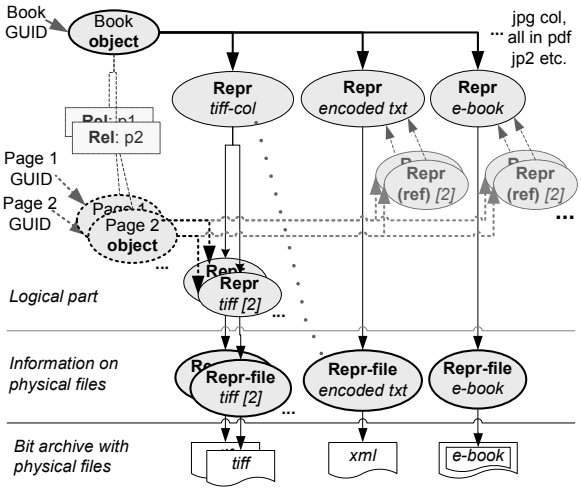


Figure 3. Modelling of a book object.

The broken lines and shapes in Figure 3 indicate that they are optional. The direction of arrows is not meant as a requirement of a concrete implementation, but an indication of the minimum information. This means, that a concrete implementation may have e.g. hasChild and isChild relations, although the arrow points one way in the model. The book object and the first layer of representations correspond to corresponding entities in Figure 2. In the *logical part* there are illustrated other levels of representations in form of pages. The part with *Information on physical files* is entities which make the link between the logical part and the physical files by referencing storage identification. The *Bit archive with physical files* is the storage, which possibly will be different according to the preservation level of the stored files. The dotted line from the *tiff-col* representation to the *xml* file for the encoded txt indicates that in the ADL case the order of elements in the collection is defined in the *xml* file. Note that the order representation could also be represented and preserved in a separate METS file, or it could be relation information metadata stored via a *Repr-file*. In the ADL case, this would mean that the current representation must be converted.

In the example there are page objects, which can relate to page representations for collection of TIFF files (*tiff-col*). However, there is no direct page

representation for *encoded txt* or *e-book*, therefore, if desired, corresponding page representations in these book representations need to be made via references into the book representation. Note that such references may not make sense for all representations, e.g. the *e-book*.

It gets even more complicated when we want to model objects that represent a literary work. A literary work can be a poem, which starts mid-page x and ends somewhere in the first half of page y. Or a literary work can be a novel that spans over two volumes (book items). This means that a literary work can be defined at different levels in the logical part of the model.

Referencing into an object means addressing a part of an object. This reference mechanism should be transformable between different representations of the object, in order to ease the work of preserving the references in different preservation and dissemination forms, e.g. for migrations. References into objects are tricky. Normally, we would think of references based on atoms like a pixel in an image, a character in an ASCII text or time past in a soundtrack. However, a pixel may get another meaning in a migration. A character or its context may be changed due to corrections in the OCR of an encoded text. Furthermore in our example a pixel will have to refer to a page image in a book, which has a challenge related to the page numbering. Another challenge is that page numbers will not be part of e.g. an e-book representation, they will have a different meaning in a representation for a mobile device, and should have a different interpretation in e.g. a voice representation. If we consider encoding mechanisms e.g. using Xlink[14] this will again need consideration on how encoding is represented, updated and related to the different representations. Furthermore, in the ADL marginal notes example, the position reference of marginal notes was the most precise.

At a starting point, we will aim at a general reference mechanism which can be translated via relations between different representations, being aware that references like e.g. page numbers will not make sense in all representations. Similar referencing considerations will need to be taken for other formats such as sound, images and maps. In the future there will be an increasing demand for representations into objects, for example annotations added via social network communities. Such examples already exist, for example, for maps[15].

Part of referencing is also how we address objects or parts of objects with identifiers. Seen from a preservation perspective, identification of an object must be unique and persistent during time. Any semantics inserted into identifiers may confuse future uses such as e.g. a format extension or structure information which does not exist in the future. An example of a semantic free persistent identifier is Universally Unique Identifier (UUID)[16]. Identification of objects includes considerations on an object definition, in the sense that the object is addressable by the identifier in the future.

A choice must be made on how an object representation is identified in the future. For example, new versions of an object may occur in form of updates with added pages. Likewise for ongoing research reports, there may be several versions of a research report. The model does support creation of new versions, since adding of extra pages can be implemented by creation of a new *tiff-col* with a version relation to the existing *tiff-col*. Additionally a new representation would have to be created for related representation, e.g. for the *encoded txt*.

Many-to-many relations can be expressed in the model on the representation level, e.g. from a *tiff-col* to an *e-book*. When doing a many-to-many migration, the preservation metadata must include details of relations on the digital objects level. Many-to-many relations may also be needed in connection with reference translation between two representations, as described in the page reference example for Figure 3.

Annotations and information from social networking can be included in different ways depending on the type of information. Examples are; OCR corrections, part of speech annotations, relations to different material, or comments on author or text.

4.1.3. Consequences for ADL Data

As we have seen, the suggested logical object model can include special cases of the old ADL material, thus this data will be able to be reused. However, there will be a need for transformation of the data, which includes a risk of loosing data. Firstly, all page references must meet final identifier standards. Secondly, we may decide to have the structure of TIFF pages separate from the encoded text, for example in a METS file. A reason for this would be to have a less complex single representation of the preserved TIFF representation.

4.2. Possible Implementation

At KB we have reached the conclusion that community around preservation and dissemination is of great importance when deciding on the implementation of a DOMS. Another high priority is to have a system with high modularity and exchangeable components, where especially preservation issues must be system independent. Lastly, a high priority is to have a system with a homogenous treatment of similar materials.

There are many both national and university libraries that face the same challenges[17]. As this research also points out, we live in a time of rapidly changing demands for what a DOMS must cover. Not all problems can be solved at once, therefore there will be different priorities, e.g. due to different focus on

[14] XML Linking Language (X-LINK). 2001. Version 1.0.
[15] Google maps, see http://maps.google.com
[16] UUID, see http://tools.ietf.org/html/rfc4122

[17] Several examples can be found e.g. in OR proceedings, for example the Mounting Books Project described on http://smartech.gatech.edu/handle/1853/28425

different materials. Thus at present no system exists which can cover all the challenges to come in the next decade. There will however be a community that faces similar challenges, and has varying overlap of priorities for implementations.

Fedora commons[18] in particular has evolved into such a community, although there are different Fedora-based applications[19] like eSciDoc, Hydra, Islandora. Fedora has an advantage in being highly flexible with regard to how the data is modelled. A disadvantage, as well as a consequence of this flexibility, is that Fedora is far from a DOMS in the sense of being an off the shelf product. Furthermore, the Fedora-based applications are primarily focussed on dissemination aspects. Yet the Fedora case seems the best alternative to meet requirements of community and ability to model data in ways that complies with the logical object model.

The flexibility in Fedora opens many ways to make a solution that respects the logical object model, e.g. by using Fedora objects solely, or by encapsulating some of the modelling aspects in use of e.g. METS. This must however be done with care[20].

High modularity and exchangeable components are important for survival of the system, in which possibilities for renewal, enhancement and maintenance of the system are vital in order to meet new demands as a consequence of new technologies for formats and dissemination. The modularity requirement is also met by most of the Fedora initiatives. The Hydra initiative meets it, even to the extent that Fedora may be exchanged with a system offering similar functionality.

A system related requirement that of the possibility for different data to be stored under different confidentiality and bit safety levels. Although it is not part of Fedora, it is possible to implement this via workflows that handle insurance of storage in differentiated ways, and through implementation of access layer respecting confidentiality aspects.

The DOMS will end up as a system where ADL will be included as a special collection, possibly with separate web interface for ADL branding. Today there exist many different small applications like ADL, which all are part of dissemination from KB, but based on different frameworks. An example is www.tidsskrift.dk which disseminates digitised journal material produced with METAe[21] into a different format and using a different navigation than ADL. However, the cost of maintaining the different applications continues to increase. Therefore ADL and similar applications will be transformed into an integrated DOMS where preservation and dissemination aspects are treated jointly. This will be in line with requirements related to homogenous user interface for dissemination and ability to integrate with other systems.

5. DISCUSSION

The ADL case study represents relatively simple cases of material. We have argued that parallels can be drawn to other materials such as images and sound. There will, however, be other characteristics for other digital materials, which need to be investigated further.

There is still a challenge to settle on a general mechanism for proper referencing into objects. We may end up with different referencing mechanisms for different types of object representations. The selected mechanism must be taken into account in migrations, since inaccuracies in migrations can mean inaccuracies in migrated reference. In any case it may be hard to foresee the endurance of strategies for referencing.

Another related question is how to handle deletion of older versions or representation. Especially if references into objects rely on special representations (like positions) then the migration must include migration of similar referencing mechanism.

Having different representations of encodings in preservation and dissemination will add sources of error. This is a balance needing risk assessment and prioritising between meeting different requirements.

There are areas of the model that are not fully described as for example how to document relations between different representations. At this stage it is not necessary to make these processes and entities explicit, but they will have to be explicit in an implementation. As for any part of the data, the bit preservation level of the descriptions must be classified and effectuated.

Another area is versions contra representations. It is not a computer scientific question whether a new edition of a book is a new version with a new object identifier, or whether it is a new representation of an existing one.

6. CONCLUSION

We have argued that demands on preservation are closely related to demands on dissemination in a library context. Dissemination has many dynamic aspects and preservation tends to aim at static aspects, focus and goals differ, and thus demands on both preservation and dissemination will add complexity when viewed jointly.

We have presented a logical object model for normalised data that can meet the preservation requirement, including dissemination considerations and future requirements for new types of representation and information from social networking.

The hypothesis that we can use old digitised data in a normalised form will hold as long as the material is transformed, which is plausible, but does also involve risk of losing data.

The next step is to update the preservation strategy according to the findings, and to develop a DOMS for

[18] http://www.fedora-commons.org/

[19] http://www.fedora-commons.org/confluence/display/FCR30/Getting+Started+with+Fedora#GettingStartedwithFedora-applications

[20] See e.g. OR 2009 contribution about Fedora 3.0 and METS on http://smartech.gatech.edu/handle/1853/28470

[21] See http://meta-e.aib.uni-linz.ac.at/

all digital materials in the library. This will include more thorough analysis of the challenge to reference into object and settle for a final implementation.

7. REFERENCES

[1] Holley, R. "Many Hands Make Light Work: Public Collaborative OCR Text Correction in Australian Historic Newspapers". *Technical Report from National Library of Australia*, Australia, 2009.

[2] Kejser, U.B. "Preservation copying of endangered historic negative collections" *Proceedings of the IS&T Archiving Conference*, Bern, Switzerland, 2008.

[3] Riva, P. "Functional requirements for bibliographic records: Introducing the Functional Requirements for Bibliographic Records and related IFLA developments" *Bulletin of the American Society for Information Science and Technology vol. 33 issue 6*, 2008.

[4] Sperberg-McQueen, C. M., Huitfeldt, C., "GODDAG: A Data Structure for Overlapping Hierarchies" *Lecture Notes in Computer Science, vol. 2023/2004***,** Berlin, Germany, 2004.

[5] Zierau, E., Jensen, C. "Preservation of Digitised Books in a Library Context". *Proceedings of the International Conference on Preservation of Digital Objects*, Vienna, Austria, 2010.

[6] Zierau, E., Johansen, A.S., "Archive Design Based on Planets Inspired Logical Object Model". *Proceedings of European Conference on Research and Advanced Technology for Digital Libraries*, Aarhus, Denmark, 2008.

[7] Zierau, E., Kejser, U.B. "Cross Institutional Cooperation on a Shared Bit Repository". *Proceedings of the International Conference on Digital Libraries*, New Delhi, India, 2010.

CAPTURING AND REPLAYING STREAMING MEDIA IN A WEB ARCHIVE – A BRITISH LIBRARY CASE STUDY

Helen Hockx-Yu **Lewis Crawford** **Roger Coram** **Stephen Johnson**

The British Library
96 Euston Road
London NW1 2DB

ABSTRACT

A prerequisite for digital preservation is to be able to capture and retain the content which is considered worth preserving. This has been a significant challenge for web archiving, especially for websites with embedded streaming media content, which cannot be copied via a simple HTTP request to a URL. This paper describes the approach taken by the British Library in capturing and replaying streaming media in a web archive. A working system is now in place which will lead to the development of more generic tools and workflows, contributing to addressing a common challenge for the web archiving community. The British Library recently archived a large scale public arts project website, http://www.oneandother.co.uk, which contains 2,400 hours of flash videos, streamed over Real Time Messaging Protocol (RTMP). The case study also presents an overview of the non-technical issues relevant to archiving this high-profile website.

1. INTRODUCTION

The web has become an increasingly important information resource for research and learning. However, the web is also ephemeral; websites disappear regularly. If not archived for long-term preservation, valuable web resources could be lost forever.

National libraries and archives around the world have been archiving the web in the 1990s. The Legal Deposit Framework of many countries now also includes the free web, with the national libraries carrying out periodical crawls of the respective national domains to capture and preserve a historical record of the web. A similar legislative framework exists in the UK but is yet to come into effect.

The importance of preserving web resources has been illustrated by the establishment and ongoing activities of the International Internet Preservation Consortium (IIPC), which was initiated in 2003 and currently has 38 member organisations across four continents. IIPC fosters the development and use of common tools, best practices and standards. Being brought together by common challenges, many national libraries and archives are active members of the IIPC, including the British Library.

1.1. Web Archiving at the British Library

With permissions from rights holders, the British Library has been selectively archiving UK websites since 2004. The Library has established an ongoing Web Archiving Programme to collect, make accessible and preserve web resources of scholarly and cultural importance from the UK domain. Archived websites to date are made available through the UK Web Archive, along with additional material archived by the National Library of Wales, the Joint Information Systems Committee, and the Wellcome Library. The National Library of Scotland and the National Archives have previously contributed to the Archive.

The UK Web Archive contains regular snapshots of over 8,000 websites and offers rich search functionalities including full-text, title and URL search. The archive in addition can be browsed by Title, by Subject and by Special Collection. The UK Web Archive was formally launched in February 2010, raising awareness of the need for web archiving, which has generated a great level of interest from the press as well as the general public.

Web Curator Tool (WCT), a tool developed by the British Library in collaboration with the National Library of New Zealand, is used to manage our selective archiving processes. WCT embeds the commonly used open source crawler software Heritrix, and has added functionalities to manage workflow. The Open Source Wayback Machine (OSWM) is utilised to render and provide access to archived websites.

In anticipation of the implementation of Legal Deposit for UK online publications, the British Library is also exploring the technical and curatorial challenges of archiving in future a much larger proportion of the UK domain, through periodical domain harvests.

1.2. The One and Other Project

The 4th plinth on Trafalgar Square in London, originally intended for an equestrian statue, has been empty for many years. This is now the location for specially commissioned art works. Between 6th July and 14th October 2009, the famous British artist Antony Gormley undertook a large scale public arts project, during which 2,400 participants occupied the 4th plinth for an hour each, doing whatever they chose to do. The project was intended to create a living portrait of the UK, providing an open space of possibility.

All participants, or *plinthers*, were filmed and the videos were brought together on the project's website: http://www.oneandother.co.uk. The websites received over 7 million visits during the project.

When the project ended in October 2009, the British Library was approached to archive the website. It was a matter of urgency as the project funding would only last to maintain and keep the website live for a limited period of time beyond the project, till end of December 2009 initially, and then extended to March 2010. This time restriction has played a significant role in some of our technical choices.

2. PROGRESSIVE DOWLOAD VERSUS STREAMING MEDIA

Broadly speaking there are two ways to deliver digital media over the Internet between a server and a media player (used locally by end users): **progressive download** and **streaming media**. The former is also referred to as **HTTP download** because media files are typically transferred from the server to a client using the HTTP protocol. In addition, the media files are downloaded physically onto the end users' device, buffered and stored in a temporary folder for the local media player to use for replay. With streaming, data packets are constantly transferred and replayed to the end users, at no time leaving locally a copy of the entire file, as is the case with **progressive download**. There are protocols, such as the Real Time Streaming Protocol (RTSP) and the Real Time Messaging Protocol (RTMP), which are specifically designed to support streaming media.

Because of the potential risk of piracy related to progressive download, many content owners choose to publish high-value multimedia data using streaming based solutions.

The collective term *rich media* is used in this paper to refer to **progressive download** as well as **streaming media.**

For the purpose of web archiving, web crawlers are commonly used to capture snapshots of websites. It generally starts from a list of URLs (*seeds*), visiting and downloading them, before identifying all the hyperlinks within the visited pages and recursively visiting and downloading these too.

Capturing multimedia content can be just a matter of determining URLs. If the content can be served by requesting it, as web pages, then the crawler will be able to download a copy of the file via a simple HTTP request, by going to the right URL. However, parsing arbitrary URLs is not always a simple task as the URL syntax can be stretched to address almost any type of network resource and URLs can be generated dynamically. Overly complex URL structures include numerous variables, marked by ampersands, equals signs, session or user IDs as well as referral tracking codes. In some cases, multimedia files are served or initiated by embedded web applications which retrieve data from the server in the background, without explicitly locating the files in the HTML.

When streaming is not via HTTP, but proprietary protocols such as RTMP developed by Adobe Systems, it is even more difficult to capture and replay the multimedia content as this requires an understanding of the implementation of the particular protocol.

3. ARCHIVING RICH MEDIA

A prerequisite for digital preservation is to be able to capture and retain the content which is considered worth preserving. This has been a significant challenge for web archiving, especially for websites with embedded streaming media content, which often cannot be copied via a simple HTTP request to a URL.

The tools currently used by the British Library, and many other national libraries and archives, do not yet have the capability of capturing and playing back streaming media content embedded in archived websites. Heritrix, the crawler software, can only capture data delivered over HTTP and/or FTP. In addition, the OSWM does not have any streaming capability.

Many organisations engaged with web archiving have long recognised the need for a solution to dealing with rich media. The Internet Archive, the Institut National de l'Audiovisuel (INA) and the European Archive for example have been actively carrying out research and developing projects to address the problem. The effort focuses on adding capabilities to crawlers for them to be able to interpret complex code and extract URLs for media files so that these can be captured by the crawlers through HTTP requests. A problem with this is that the exercise of URL parsing needs to be frequently repeated as sites such as YouTube constantly change the way of publishing videos to prevent direct downloads. In addition, the replay aspects of the captured media have pretty much been left to the capability of the browsers, or occasionally solutions developed specifically for individual media player applications.

Adding capability of capturing and replaying rich media in web archives is a key area of work for the IIPC.

4. CAPTURING AND REPLYING PLINTHER VIDEOS

The One and Other website contains 2,400 hours of video in .flv format, approximately 1TB, streamed directly over RTMP. Initial test crawls of the sites using the Web Curator Tool (essentially Heritrix) only brought back static HTML pages without the videos, which the artist and curator considered as significant and essential components of the project and the website.

As previously mentioned, the One and Other website had a planned take-down date of end December 2009, which only allowed us a couple of months to find a solution to capture the website (the take-down date of the website was later extended to end March 2010). There was additional pressure to develop an access solution too, as the plan was to invite the artist Antony Gormley to speak at the formal launch of the UK Web Archive three months later to maximise the impact of the event. The tight timescale meant that our goal was to find a working solution for an immediate problem, rather than setting out to develop a generic technical solution for the long term within that phase of the project.

4.1. Capture

Essentially a combination of a browser and a streaming media recorder was used to initiate and capture the video streams from the One and Other website. The choice of software was largely determined by its functionality being adaptable to the project at hand. Apart from test captures to check reliability and quality, the main criteria used to select a streaming media recorder included the ability to capture media steamed over RTMP, to schedule captures and the ability to import a schedule so that a degree of automation was possible. It was equally important that the chosen software's method of naming the captured files should allow easy identification of the video along with the web page it was captured from.

Based on the above criteria, we chose Jaksta as our media recorder. Jaksta can detect videos and music streamed over RTMP, using port 1935, and capture the TCP/IP packets as they are sent to the embedded flash player in the browser. Although not allowing imports, Jaksta uses a sqlLite database which gave us the opportunity to automate some parts of the scheduling.

Prior to the actual captures, a Unix shell script was used to identify pages containing video streams, which output a list of URLs of pages containing videos. Four virtual machine instances, all configured with Jaksta for capturing video and SqlLite2009 for scheduling, each based on the schedule launched Internet Explorer instances at three different URLs at a time, to initiate the video streams. It was then a matter of letting Jasksta do the job of capturing the videos. The scheduling, also inserted using a Unix shell script, was set at 90 minutes intervals. We knew in advance that each video was approximately an hour long, so this was the metric used as a static variable to create scheduling.

Once completed the captured video was saved onto local disk. Jaksta uses the URL query as a naming convention when possible, which suited us and allowed easy identification of the link between the video file and the web page which it was embedded in and captured from.

The method described above was used to capture the video content from the One and Other website. File size was an immediate attribute used to monitor the capturing process because all the videos are of similar length, and significant variance in file size was an indication of error. File size in itself, however, cannot determine definitively if the full was captured. A shortcoming of Jaksta was that that it did not recognise or report when the full video was not captured. The videos were also spot-checked by viewing them, validated using the FLVCheck tool (by Adobe), and where required and possible, repaired using FFmpeg.

A second attempt was made to re-capture a portion of the videos which appeared shorter in length but this made no difference, which made us suspect that the error may be inherent to the video files themselves. This was confirmed when SkyArts, who sponsored the One and Other Project, later provided us with the original video files on a disk which unfortunately contained the same errors. The errors were believed to be caused by the videos in question not being recorded as one file, resulting in a mismatch between the metadata layer and the content layer. As a result, these videos have been curtailed in the web archive and cannot replay to the full length. SkyArts is currently looking to fix these videos.

4.2. Replay

Capturing the videos only completes half of the job. In order to provide access to the archive version of the One and Other website, we also needed a solution to play back the videos, as part of the end use interface of the UK Web Archive.

When granting a licence to the British Library, SkyArts explicitly required that the video content may only be streamed to the archive users, having in place the copy protection equivalent to that applied to the original website. This requirement eliminated the possibility of implementing any solution based on progressive download.

Two open-source software tools have been used to stream and replay the videos. Red5, a Java media server, was chosen as our streaming mechanism. In addition to the base streaming server, Red5 requires an application to access and serve the media. Several demo

applications can be installed by default and the 'oflaDemo' application, designed simply to serve from a flat file system, was adequate to serve this purpose. For the client side, Flowplayer has been selected as the video player, used to play back the streamed Flash videos.

In order to replace the original flash objects and to reference the local videos, a modification has been made to our Wayback timeline implementation, which is a banner inside rendered HTML pages inserted by the OSWM, allowing users to navigate between individual archived versions of the current page. A few lines of JavaScript has been added to firstly, if not already defined, reference the flowplayer() function by calling the flowplayer-*.min.js file. The window.onload function has then been amended to load a Javascript file with the same name as the original domain from the Flowplayer location (i.e. http://...wayback/*/http://www.oneandother.co.uk/ will load www.oneandother.co.uk.js). This contains a single function - streamVideo() - which does two things:

1. Replace any existing Flash elements with an object of equal dimensions.
2. Call the now-defined flowplayer() function, passing in (among other things) the name of the video file, derived from the plinther's name, and the name of the above, new object.

The One and Other website is no longer live on the web since 31 March 2010. The domain name oneandother.co.uk now redirects directly to the archival version in the UK Web Archive: http://www.webarchive.org.uk/ukwa/target/32145446/

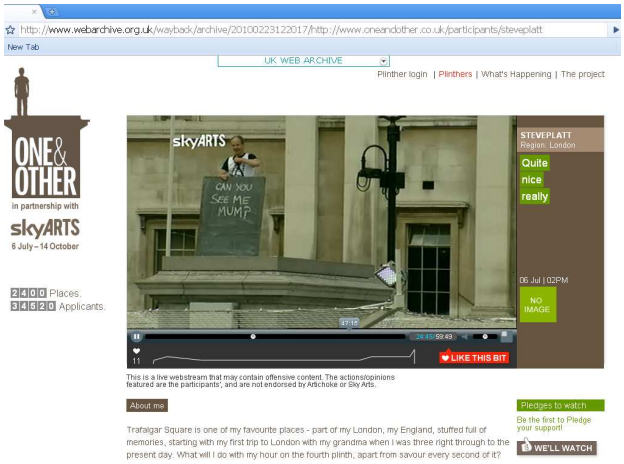


Figure 1. A screenshot of the archived One and Other page

5. NOT JUST TECHNICAL CHALLENGES

One and Other was the most talked-about arts project in the UK in 2009 and it caught the media's attention from the very beginning. Archiving such a high profile website involving multiple stakeholders meant there were also legal, curatorial and communication challenges which required the team's extensive attention.

Even before any technical solution was attempted or experimented, the media had already reported that the British Library would archive the One and Other website and preserve it in perpetuity. Publicity, when well managed, can however help the cause of web archiving. Antony Gormley was invited to the formal launch of the UK Web Archive at the end of February 2010, who spoke positively about working with the British Library. This has helped generate positive publicity and illustrate the importance of web archiving.

The One and Other project had many stakeholders, including the artist, the sponsor, the producer, the technology provider and the 2,400 participants. Intensive interaction with the stakeholders took place to coordinate and communicate the archiving process. It is not always possible to balance the interests and expectations of all the stakeholders. There is generally an expectation for an archived website to behave exactly the same as the live website. For some plinthers, it is difficult to appreciate the concept of an archival website and understand why message boards and discussion forums no longer work.

The British Library has a standard licence which website owners sign to grant us permissions to harvest, provide public access to and preserve their websites in the web archive. A customised licence had to be developed specifically for the One and Other web site, introducing additional terms and conditions and specifying in detail the involved parties' obligations.

There had also been a couple of occasions in which a plinther or a third party had requested that certain pages of the website to be taken down. Delay in taking actions or non-compliance could have potentially resulted in legal proceedings. These occurred when the live website still existed and were dealt with by the sponsor and the artist directly. Although all that the Library was required to do was to recapture a "cleaned" version of the website, such situations, however, do raise interesting and significant legal and curatorial questions. When the live website no longer exists, the Library would be seen as republishing the website by providing public access to its archival version. This in itself will transfer certain legal risks from the original publisher to the Library.

6. NEXT STEPS AND CONCLUSION

The British Library has invested considerable resources in archiving the One and Other website and successfully implemented a solution within the required, extremely tight timescale. The addition of the One and Other website to the UK Web Archive has helped raised the profile and awareness of web archiving. Our approach to capturing and replaying streaming media seems to be the

only way at the moment to capture the video streams as they are not available via standard (HTTP) protocols. It is the first practical streaming media implementation within the international web archiving community and provided us with valuable hands-on experience which will lead to more generic solutions.

A main issue with the solution described in the case study is that it sat outside the operational workflow. The video files for example were not stored as ARC files with the rest of the web archive but streaming from a separate video server in the native format. This introduces data management complexity and potential digital preservation risks.

It is desirable to build streaming media capability into the current web archiving tools commonly used by the national libraries and web archives. Alternatively we could extend the open source tools we used for them to interpret archived websites. We are pleased to report that in the subsequent months following the project, we carried out further development work on Red5 which can now stream from non-compressed WARC files.

The LiWA project, funded by the European Commission, has recently released a rich media capture plug-in for Heritrix which aims to enhance its capturing capabilities to include HTTP downloads as well as streaming media. It is still an experimental version of the software but nevertheless shows potential of adding rich media capturing capability to Heritrix.

The advent of HTML5 in addition seems to offer the most effective solution to replaying HTTP media in a web archive. The introduction of the <video> tag explicitly marks up the content which means video can be streamed over HTTP and replayed directly by the browser without the necessity of additional applications.

The recent technological developments are encouraging and it is not unrealistic to expect in the foreseeable future a solution to capturing and replying rich media in web archives. In parallel, the web archiving community perhaps should also consider approaching major rich media publishers (e.g. YouTube) to achieve collaborative arrangement so that more focused solutions can be developed at the API level.

7. REFERENCES

[1] The British Library Web Archiving Programme: http://www.bl.uk/aboutus/stratpolprog/digi/webarch/

[2] European Archive: http://www.europarchive.org/

[3] FFmpeg: http://www.ffmpeg.org/

[4] FlowPlayer: http://flowplayer.org/

[5] FlVCheck Tool: http://www.adobe.com/livedocs/flashmediaserver/3.0/docs/help.html?content=06_admintasks_11.html

[6] Heritrix: http://crawler.archive.org/

[7] International Intenert Preservation Consortium (IIPC): http://netpreserve.org/about/index.php

[8] Internet Archive: http://www.archive.org/

[9] Institut National de l'Audiovisuel (INA) : http://www.ina.fr/

[10] Jaksta: http://www.jaksta.com/

[11] LiWA project: http://www.liwa-project.eu/

[12] LiWA rich media capture module for Heritrix: http://code.google.com/p/liwa-technologies/wiki/RichMediaCapture

[13] Open Source Wayback Machine: http://archive-access.sourceforge.net/projects/wayback/

[14] Red5: http://osflash.org/red5

[15] The UK Web Archive: http://www. webarchive. org.uk/.

[16] The Web Curator Tool: http://webcurator. sourceforge. net/

ADDING NEW CONTENT TYPES TO A LARGE-SCALE SHARED DIGITAL REPOSITORY

Shane Beers
University of Michigan
Preservation and Conservation
3215A Buhr Building
837 Greene St
Ann Arbor, MI 48109-3209

Jeremy York
University of Michigan
300 Hatcher Graduate Library North
920 N. University Avenue
Ann Arbor, MI 48109-1190

Andrew Mardesich
California Digital Library
University of California,
Office of the President
Oakland, CA 94612

ABSTRACT

As a digital repository for the nation's great research libraries, HathiTrust brings together the immense collections of partner institution. Initially, the Submission Information Packages (SIPs) deposited into HathiTrust were extremely uniform, being constituted primarily of books digitized by Google. HathiTrust's ingest validation processes were correspondingly highly regular, designed to ensure that these SIPs met agreed-upon qualities and specifications. As HathiTrust has expanded to include materials digitized from other sources, SIPs have become more varied in their content and specifications, introducing the need to make adjustments to ingest and validation routines. One of the primary sources of new SIPs is the Internet Archive, which has digitized a large number of public domain materials owned by HathiTrust partners.

Many of the technical, structural, and descriptive characteristics of materials digitized by the Internet Archive did not match previously developed standards for materials in HathiTrust. A variety of solutions were developed to transform these materials into HathiTrust-compatible AIPs and ingest them into the repository. The process of developing these solutions provides an example to other organizations that would like to add new types of materials to their repository, but are uncertain of the issues that may arise, or how these issues can be addressed.

1. INTRODUCTION

As a digital repository for the nation's great research libraries, HathiTrust brings together the immense collections of partner institutions.

HathiTrust strives to conform to the characteristics of a Trustworthy Digital Repository [1], and a significant amount of work has gone into developing ingest functionalities that analyze SIPs to determine whether they meet a number of standards. The standards include the technical aspects of the digital image files in a SIP (such as resolution, well-formedness, compression type, color and bit depth), descriptive elements of the SIP (including PREMIS preservation metadata and image header metadata), and structural metadata that explain what the digital image files represent and allow software tools to display the images correctly.

The majority of SIPs being deposited into HathiTrust initially were books that had been digitized by Google, Inc. The specifications Google uses in its digitization package were worked out collaboratively with Google library partners, resulting in a tightly controlled technical and descriptive SIP. The validation environment employed in HathiTrust was developed around the ingest of these materials. For some time, this ingest process has worked well in verifying SIPs against set standards, allowing content into the repository when compliant, and reporting when something failed.

However, the scope of digitization at HathiTrust institutions is much broader than Google digitization alone, and one of the partners' initial goals was to accommodate the outputs of the variety of digitization initiatives they had undertaken in a single repository. Because a number of partner institutions have had materials digitized by the Internet Archive (IA), expanding the capabilities of HathiTrust to preserve and provide access to these materials was a logical and highly desirable direction to pursue.

In the summer of 2009, the University of California (UC) was poised to deposit an initial set of nearly 100,000 IA-digitized volumes into HathiTrust. Talks were initiated between staff members at California Digital Library (CDL) and the University of Michigan on how to accommodate ingest of this content, and in the fall of 2009 a core team from the two institutions was formed to work out the details of ingest. The team worked over a period of nine months to develop specifications and routines for ingest of IA-digitized volumes generally, and HathiTrust began downloading UC content from IA in April 2010. This paper describes the issues the team encountered during this process and the solutions implemented to create a sustainable large-scale process for ingesting this new content.

2. ISSUES FACED

While partners wanted content digitized by IA to be preserved in HathiTrust, many of the technical, structural, and descriptive characteristics of this content did not match the previously developed standards for materials in the repository. The following are some of the issues the ingest team faced:

Issues related to IA Identifiers:

- The characteristics of primary identifiers would be problematic in HathiTrust systems.
- Filenames differed from HathiTrust conventions.

Resulting Questions:

- What can be used as a primary identifier?
- How will this decision be made?
- What accommodation, if any, will be needed for the different file-naming scheme.

Issues related to IA File Types and Metadata:

- Both raw original and edited page images were present.
- Metadata was located in a number of separate files, and metadata files were not present in a consistent manner between packages. Additionally, none used any obvious schema.
- Some files that the Internet Archive maintained were of undetermined value for preservation; a "preferred" package needed to be identified.

Resulting Questions:

- Is it prudent to preserve raw originals and make them accessible?
- What information captured by IA meets the requirements in the existing HathiTrust AIP specification?
- How do we deal with missing metadata and metadata that does not meet the requirements (e.g. differently formatted dates, invalid MARCXML, etc)?
- What should be done with the metadata in the IA SIP that is not part of the current HathiTrust METS profile?
- What PREMIS syntax do we use to properly record the transformations made to the SIP?

Issues related to IA Page Captures:

- Captured images did not always represent actual page data (e.g. captures of the cradle, tissue papers, and scanning targets).
- Some page types indicated a lack of label authority control (e.g., "Title Page" and "Title" being used to represent the same type of page) or contained errors (e.g. "Norma" instead of "Normal").
- Some required technical and descriptive metadata elements were missing from the image file headers.

Resulting Questions:

- How do we manage structural issues, such as erroneous page types and scanned pages that should not be displayed?
- How do we map the IA page tags to the standard HathiTrust values?
- If image header information is missing, can it be safely and reliably derived from the image data or assumed to be a standard value?

These separate questions led to two overarching issues for the team to address: what transformations would be needed to create HathiTrust-compatible AIPs from IA SIPs, and in what ways could the ingest verification process be modified to accommodate IA-digitized content, but still maintain a high degree of consistency and corresponding reliability for preservation across the repository?

3. SOLUTIONS DEVELOPED

To address these issues, the team of staff members from CDL and Michigan met over a period of months, consulting both with HathiTrust partners and non-partners who had digitized content with IA, to overcome the technical hurdles to ingest. Ensuring the long-term preservation of the digital materials was the highest priority in the development of strategies, with the simultaneous desire to ensure access to the ingested objects. Successful alignment of the Internet Archive SIP to HathiTrust standards required team members to balance the following specific objectives:

- retain components of the SIP that were most useful for preservation and access purposes
- create the most efficient ingest package in terms of size and number of component parts
- maintain functional consistency across the repository
- develop procedures and policies that could be generalized to future types of new content

3.1. IA Identifier

One of the first questions the ingest team encountered was whether to continue using IA's primary identifier for volumes as their identifier in HathiTrust. Tagged as "<identifer>" in the object's meta.xml file, the IA ID is used in the names of all files associated with a given object, and is also embedded in the object's URL hosted by IA. While the IA ID works well for Internet Archive's own purposes, the ingest team found it could not easily be integrated into the HathiTrust environment:

- While the majority of IA IDs contained only lowercase characters, several were found with uppercase characters. IDs need to function in case-insensitive contexts in HathiTrust, and team

members found that IA IDs were not necessarily unique when lowercased.

- IA IDs had no distinct length. A set of identifiers representing 190,000 objects averaged 24 characters long; a small proportion of this set was found with over 30 characters, and some over 40 characters. Lengthy identifiers would strain the HathiTrust catalog, as well as the pairtree implemented directory structure.
- IA IDs contained embedded semantics: author, title, volume, and scanning facility. Semantics put unnecessary weight on an identifier when the goal is long term preservation. For instance, a string of letters could carry a different unintended meaning in some other time or place.

Fortunately, through collaboration with CDL in developing its processes, IA also generated a NOID (nice opaque ID) for each object [2], prefixed with "ark:/" and written to the meta.xml file. Ultimately, the NOID was chosen to be the primary identifier within the HathiTrust AIP. The original IA ID was retained in the METS for purposes of posterity, but is not used to access the object.

The NOID identification scheme was chosen as a primary identifier for this new ingest type because: 1) The NOID was already embedded in the object's metadata record. 2) The NOID was created directly by the digitizing agent instead of by a receiving institution. 3) Being an opaque and short identifier, a NOID is unique across all providers 4) NOID supports the ARK (Archival Resource Key) scheme [3], which – although not fully implemented in the current HathiTrust instance – dictates a tight binding between an ARK URL (a combination of a NOID with some name mapping authority) and its metadata.

Ideally, an identifier should correspond to the identifier in use for the physical object, embodied, for instance, in a scannable barcode. Although the identifier scheme decided upon for these books in question did not involve a tight binding between identifier and object, the team believed it arrived at a durable compromise.

3.2. File Types and Metadata

One of the most significant issues faced was the difference between the structure and content of the Internet Archive book packages and packages already preserved in HathiTrust. The Internet Archive scanning process creates a variety of files in different formats, and generates significantly different metadata than those produced through Google process, for example, or other locally-digitized content contained in HathiTrust. Files chosen for long-term preservation from IA had to be carefully selected, with attention to both near- and long-term utility and viability.

The ingest team decided to select certain files from the IA SIP for preservation and exclude others. Any file that contained information determined to be valuable was kept. These included primary images in the JPEG2000 format, information describing how raw images were captured and modified, MARC cataloging information, and OCR data. Any file that could be re-created from the preserved content was excluded, such as a PDF version of the book, .GIF images, .DJVU files, and Dublin Core metadata. After some debate, the raw, uncropped page captures were not preserved for several reasons: their value above that of the cropped images was unclear; they required an additional 1.5 to 1.75 times more storage space than the cropped page images, which were already of significant size; and they would need to be processed to be used in the same manner as the cropped images, which HathiTrust did not support.

A set of pertinent files were thus selected for inclusion in the HathiTrust AIP. However, further analysis of IA SIPs found that not all of these files were present consistently in the SIPs. The files were therefore further divided into "core package" files that would be required in each IA SIP and "non-core package" files that were highly desired, but determined in the end to be optional. The package designations were based on the ingest team's determination of which files were most valuable for preservation and access purposes. The core package contains the image files, OCR data, and the core descriptive metadata and scanning process metadata. The non-core package contains file checksum data, and potentially useful but non-essential scanning process metadata. The team decided to use PREMIS metadata [4] to document any non-core package files that were missing from an SIP. If core package files were missing, the volume would not be ingested.

3.2.1. IA METS Document

Perhaps the most interesting decision made in the process of accommodating the IA SIPs was one to create a separate METS file in the AIP to store the information contained within the metadata files retained from IA, and then discard the original IA metadata files themselves. This was consistent with the existing practice for Google packages, where a Google-produced METS file is stored in the HathiTrust AIP in addition to a functional METS file (the HathiTrust METS) created by HathiTrust for its own use in the repository. A single METS container for information from the IA files would allow the team to save valuable information in a way that simplified management of files and maintained consistency in the repository, both in the overall package specification and in the HathiTrust METS. The HathiTrust METS would therefore not need to be modified to accommodate these new elements. Instead, including some base information from the IA METS file (such as creation date, as is the practice for Google-provided METS file), the HathiTrust METS could be a record primarily of actions and events occurring in relation to an object after its ingestion into the

repository, while the IA (or generically, digitization source METS), could function as the record of the digital object prior to ingest. Though previously a peculiarity of Google-digitized content, the idea of combining all information about digital materials prior to ingest in a single file took hold in the IA ingest process, and has become integral to strategies for ingesting content from a variety of digitization sources.

The IA METS is built by parsing the separate metadata files inside the IA SIP and copying their contents into a METS file similar to the one that is part of each HathiTrust AIP. This takes place during a pre-ingest phase, which the team developed to affect all modifications relating to metadata and content in the IA SIP (e.g., image headers), prior to final validation and ingest. The IA METS is similar in format to the HathiTrust METS that is part of each AIP. Most of the IA METS is boilerplate structure, filled in with information downloaded from the IA book package or the objects as they are processed for HathiTrust compatibility. The information in the IA METS includes MARC XML, descriptive metadata, OCR information, and metadata about the scanning process – all of which were part of the IA SIP but not necessarily appropriate for inclusion in the HathiTrust METS.

3.3. PREMIS Events

The transformations and processes that occur during the pre-ingest transformation are documented in the IA METS using PREMIS metadata in order to maintain the digital provenance record, with the goal of providing additional trustworthiness. The decision was made to employ PREMIS 2.0, as opposed to the PREMIS 1.0 used in Google- and other partner-digitized AIPs, because it allowed for new preservation elements, and repository-wide plans included transitioning all content to PREMIS 2.0. The transformation events include processes such as MD5 validation, IA SIP inspection, image header modification, file renaming, OCR splitting, IA METS creation, and final validation. PREMIS is utilized to document the processes and actions performed, the institution that performed it, and the software tools employed.

3.4. Image Headers

Addressing missing image header metadata was somewhat complex. HathiTrust requires JPEG2000 files to have technical and descriptive metadata in the XMP box, but this information was not always present in the IA images. The ingest team decided to use ExifTool to modify and/or populate metadata in the image headers if it could be reliably derived or taken from metadata provided outside the headers. Some of this metadata, such as TIFF:SamplesPerPixel and TIFF:PhotometricInterpretation, could be derived from the bitstream using JHOVE. TIFF:Orientation was assumed to be 1 (which indicates a horizontal, or normal, orientation), as images were captured in the orientation in which it should be displayed.. Some elements were able to be copied from the JPEG2000 metadata elements such as the image width and height. One of the more difficult issues faced was missing JPEG2000 resolution information. Here the team decided to determine the resolution value from data found in the file header in the JPEG2000: CaptureResolution and CaptureResolutionUnit fields. If this was not present the resolution was determined by using information captured in the IA metadata files, which appeared to match the resolution metadata in the header when present.

3.5. Page Types

There were a number of issues with individual page captures in the IA SIPs that needed to be resolved. Among the page captures were images of the scanning stand, scan targets, tissue pages, and miscellaneous pages that were tagged as “delete”. A of a lack of documentation of this portion of the digitization process, it required the ingest team to deduce what was meant by some of the labels (e.g., identifying tissue pages, blank pages, title pages, tables of contents, etc.). Even after these variations were clarified and misspellings were normalized, these labels did not always neatly fit into the standard array used in the HathiTrust AIP. In the end, original IA page type values were stored in the IA METS and normalized to HathiTrust values during the creation of the HathiTrust METS. Where applicable, some page type values were incorporated as additions into the standard HathiTrust schema for labeling pages. While it would not have been a burden to accommodate IA labels in the HathiTrust access system (where they are used to browse content) instead of normalizing them, the team determined that asking downstream users of HathiTrust content to analyze the different DIPs (Dissemination Information Packages) to understand multiple labeling schemes would unduly inhibit use of the content.

4. FINDINGS AND CONCLUSION

Much was learned in the process of developing successful and appropriate methods for ingesting IA-digitized materials into HathiTrust:

- Documentation of process is essential to downstream uses of content. Significant time was spent by the ingest team in analysis and interpretation of IA processes and digitized content because documentation was not available. In some cases such as foldout images (which are not gone into above) no special action was needed for preservation or display purposes, but

extensive investigation was required to determine that this was the case.

- File names are just names, and should not be invested with too much meaning. Much deliberation occurred around filenames but in the end the team decided to use the IA names instead of normalizing them. The issue of primary concern is that metadata exists to indicate the proper order of files, not the filenames themselves.
- In a collaboration of this size, with expertise required in so many areas, open and clear communication is the key to success. Agendas for meetings, facilitators of conversations, and individuals at each institution to coordinate efforts, talk through issues, and bring in additional team members for perspectives, insights, and expertise as needed, were essential to the success of this project.
- The trustworthiness and effectiveness of a shared repository does not rely on specifications and sound technology alone. They are based as well on the relationships of the people involved in building and sustaining the repository over time. Through the conversations and experiences working together, the teams from CDL and Michigan gained greater trust in one another, and in the methods and processes we use for getting things done. Building relationships through in-person, phone, and video conferences throughout the project helped the team accomplish its goals, and will strengthen HathiTrust in its collaborative efforts going forward.

The collaboration between the HathiTrust partners set precedent for future approaches to ingesting content from new sources. The contributions from each institution and other HathiTrust partners led to a strong shared philosophy on digital preservation and content management.

This type of experience is likely to be far more common as digital repositories seek to expand their stores of digital content to content produced by a variety of providers and partners, while simultaneously attempting to create strict validation routines and a manageable, consistently-structured store of AIPs. It is hoped that this case study will provide a model for other organizations and collaborations to follow as they expand their collections.

A large number of staff from the University of Michigan and California Digital Library contributed to the success of this project. The authors would like to acknowledge their efforts and offer thanks for their contributions to this paper.

5. REFERENCES

[1] TRAC. (2007). *Trustworthy Repositories Audit &Certification: Criteria and Checklist.* Center for Research Libraries and OCLC. http://www.crl.edu/sites/default/files/attachments/pages/trac_0.pdf

[2] NOID: Nice Opaque Identifier (Minter and Name Resolver). (2006). https://wiki.ucop.edu/display/Curation/NOID

[3] ARK: Archival Resource Key. (2008). https://wiki.ucop.edu/display/Curation/ARK

[4] PREMIS: Preservation Metadata Implementation Strategies. http://www.loc.gov/standards/premis/).

NATIONAL FILM BOARD OF CANADA DIGITIZATION PLAN – A CASE STUDY

Julie Dutrisac, Eng. **Luisa Frate, CA** **Christian Ruel**

National Film Board of Canada

3155 Côte-de-Liesse, Montreal, Quebec, Canada

ABSTRACT

The digital revolution has completely changed how audiences use and interact with audiovisual media. The National Film Board of Canada (NFB) has been preparing for this inevitable revolution for several years now, developing partnerships and carrying out research on image and sound processing, innovative transfer techniques, accessibility and distribution to facilitate the transition to digital technology. In recent years, the technical and operational infrastructure has undergone significant upheavals. And the NFB has taken up the gauntlet. The new digital reality is much more sophisticated than past technologies, but also much more open and promising. Our aim is to exploit the full potential of these new and constantly evolving technologies.

1. INTRODUCTION

In this context, the accessibility of the works the NFB produces and distributes is a major priority as well as part of our mandate.

The institution serves Canadians by making its rich collection and productions available to them when and where they want and on the platform of their choice. The NFB's digitization plan is an important step in achieving those goals.

One of the NFB's main objectives is worldwide on-line accessibility of its extensive collection, either through excerpts or full-length streams. The NFB must be able to offer a range of formats and platforms, from D- or E-Cinema in movie theatres to HD television and Internet broadcasts, and from downloads on different platforms – including mobile devices like the iPhone and the iPad – to traditional DVD or Blu-ray discs for home viewing.

The objectives of the NFB's digitization plan are to improve current and future accessibility of NFB works in digital formats, store and preserve the NFB's works on new media, and restore works that have deteriorated.

The NFB digitization plan applies primarily to finished audiovisual works. This collection, dating back to 1939, is made up of 12,963 titles on film and video in a wide range of formats, from 16mm, S16mm and 35mm to 70mm, and in a variety of magnetic, optical and digital sound formats.

Most of the NFB collection has been produced in multi-language versions (usually in French and English), and since early 1990, all films have been closed-captioned.

Faced with the challenge of digitizing its huge collection, the NFB has created workflows based on the type, physical condition and variety of source material to be processed for a given work. In addition, the technology current when the work was produced must be taken into account so that as much relevant information as possible is captured, all the while respecting the creative choices made at the time.

The NFB's plan to digitize its collection requires innovative workflows. The purpose of the workflows is to process the largest number of works possible. Specific workflows will be established to handle exceptions and other difficult cases. This approach should ensure efficiency, but also allow restoration in a way that respects the originals.

2. DIGITIZATION PLAN VISION AND TECHNICAL CONCEPTS

Our vision of the digitization and distribution of the NFB's collection may be summarized as follows: for each work in the collection, a Digital Source Master (DSM) will be created to preserve the work. Each DSM will be made up of its individual component parts in an uncompressed format: the image, sound, metadata and effects. Every segment of an NFB work (all image and sound segments, titles in all existing languages, subtitles, credits in all existing languages, closed-captioning files, etc.), found in every one of its versions, will be digitized and processed only once at a sufficiently high resolution to allow delivery in all of our distribution and access formats as well as to create a digital master for digital preservation. Damaged or deteriorated works can be digitized at a higher resolution (4K or 6K) to ensure that they can be restored and preserved. The assembly rules

followed in producing each version will be saved along with all other data that will help us understand the processing performed on each component. This content becomes the work's Digital Source Master (DSM), and all of these components will be archived. It should be noted that this practice of digitizing and processing each work's individual parts only once will make the whole process more efficient and flexible.

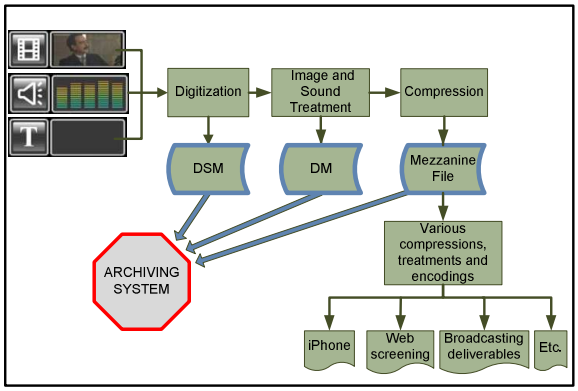


Figure 1. Major elements of the digitization plan.

Another advantage to this approach is that it reduces the volume of data archived for each work, particularly since image data files are much larger than their sound data counterparts. Furthermore, for preservation purposes, this approach allows us to select the best original source for each component. Thus, for works on film, an assessment of all available sources in each version and every format that was produced (negative, interpositive, internegative) will help determine the best source material for each segment digitization.

Producing appropriate descriptive and technical metadata for all audiovisual material and all processing involved in the material's life cycle is both attractive and essential in order to pursue fully file-based production and distribution, as well as for automation of workflows.

Metadata then needs to be closely attached to the media essence and embedded in the headers of the various files. Metadata must also be ingested into the archival system database so it can be easily searched, read and manipulated.

Metadata is the key to producing multiple versions of works from their many digital components, because it provides us with a better understanding of a wide range of parameters and characteristics.

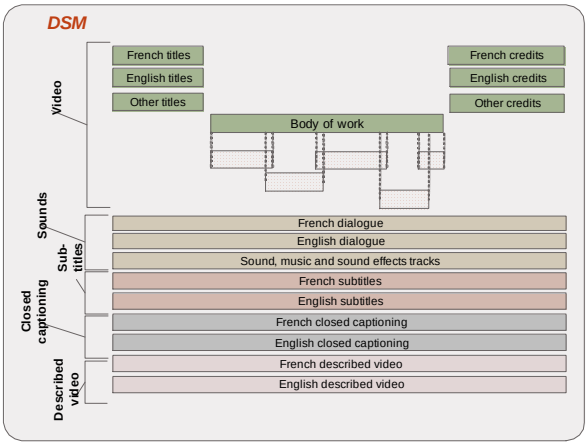


Figure 2. Digital Source Master (DSM) approach.

Once the digitization and metadata capture have been performed and checked, the completed DSM and its metadata for the work are archived. Integrity of the content is checked when accessed or ingested into the archival system database using MD5 checksums. To ensure the security of our collection, data replication is achieved using rules defining the number of duplicates to make on different media and their storage locations. These rules, based on our collection preservation strategy, rely on the state and availability of the original source elements, the availability of the playback equipment and the heritage value of a title.

At this stage, the uncompressed DSM components are processed in order to return the work to a state as close as possible to the original and to recreate the various versions of the work. Processing includes restoration, colour calibration and component synchronization. This processing will result in an uncompressed Digital Master (DM) that will be archived with all of the metadata collected at each processing step. The uncompressed and unaltered DSM will be kept for future restoring and further processing improvements as digital-image and sound-processing technology evolves.

The archiving process of the DSM generates over 400,000 metadata entries associated with 1.5 TB of content and 100,000 files for an hour-length work, with a very simple version in 2K resolution. The same figures are used for archiving the DM.

3. CREATION OF MEZZANINE FILE AND ACCESSIBILITY DELIVERABLES

At this stage, the finished, uncompressed DM contains all the unassembled segments and the various assembly lists allowing the creation of all the versions of the work. Before being archived, it will be run through a two-phase quality control check. The first phase is a fully automated process of planned and systematic verification and validation operations to ensure that all elements are up to established standards. In the second phase of the quality control workflow, all files (DPX,

Broadcast Wave, subtitle text, etc.) of the various segments are ingested with the various versions' assembly lists, and a complete viewing of the work and its various versions is done to confirm and validate all choices made. Upon approval of the DM, the unassembled segments of media essence and metadata are automatically wrapped as an MXF-AS02 mezzanine bundle. Automatic processes are also initiated to archive the DM and to archive the mezzanine file.

The use of the MXF standard with the AS02 specification for the mezzanine files offers the possibility and flexibility to manage multiple versions of a work within a single mezzanine file bundle without creating multiple copies of the media essence. The MXF file format is widely used within the industry, and supports multiple video compression standards and the management of multiple segments and their reassembly.

This compressed MXF-AS02 digital bundle, generated from the DM, is used to create the NFB's main production and distribution deliverables. Its degree of compression is determined by the types of deliverables to be produced. The mezzanine will meet our primary deliverable needs. On demand, the required deliverable will be created from the mezzanine file MXF-AS02-wrapped content by rendering the specific simple version in an MXF-AS03 delivery format.

We plan on using the mezzanine file for Internet, DVD (traditional and Blu-ray), download, mobile platform and television deliverables. For some types of deliverables, it will be necessary to return to the DM, particularly for D-Cinema or transfers back to film. Rules will be put in place to manage files based on their usage history. In addition to being adopted by the industry's major players, this approach will also meet our current and future accessibility requirements.

4. METADATA AND ARCHIVING

Long-term retention and management of digital assets is another key requirement for this digitization plan, which must contend with exponential growth in data volumes, long-term retention requirements and on-demand-deliverables mandates on a per-customer basis. When content is archived, its specific parameters must be recorded to ensure proper categorization and allow effective and efficient search and analysis.

A robust, reliable archiving solution is a key component of our media operations for the collection, preservation and distribution of our digital assets. The archiving solution needs to ensure long-term retention of the digital collection and efficient management of the exponential growth in data volumes from all new productions. It also needs to provide efficient tools for the complete asset life and allow efficient search, retrieval and cataloguing functionalities along with scalability to handle the growing data volumes required to support the digitization plan and each year's new productions.

One challenge is to change the organizational structure of information to adapt it to the new opportunities and flexibilities possible with the digital shift. Notably, we have refashioned the structure to link all information to a work in all of its possible versions instead of the usual way of addressing single titles of a work. This approach helped to achieve the digitization plan's vision of creating a single DSM per work and a seamless integration of the archiving solution with our existing asset management, which stores all descriptive metadata of each work and its specific versions.

For each digitization and processing activity, the maximum amount of metadata, within reason, is collected. All descriptive metadata resides in the asset management database and all administrative metadata (i.e., of technical and digital provenance) is kept within the archival database system, with some parts being embedded in the various files.

In order to enable the organization of our digital assets in a hierarchy tailored to our workflows, it was important that the archival solution not dictate a specific archive structure. The metadata associated with the digitization of a film was implemented from the Digital Picture Exchange (DPX) file format ANSI/SMPTE standard (268M-2003). The data model implemented is presented in figure 3.

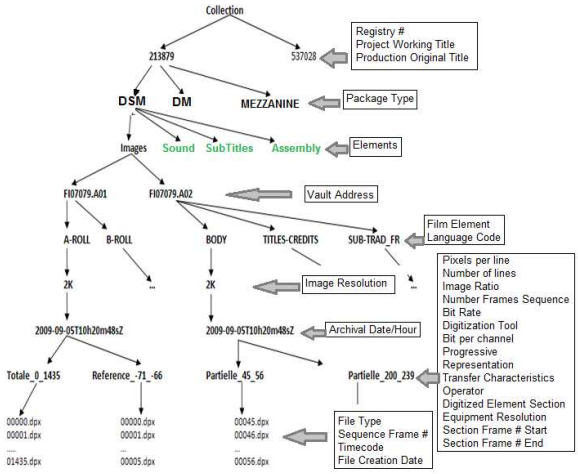


Figure 3. Film digitization data model.

For the audio workflow, metadata related to the DSM is collected for the source audio object at the recording level describing the technical attributes of all physical audio sources (physical characteristics of the original material, format of the source and specific information on transfer and creation). In each activity related to the processing of the sound components of a work, from recording to finalizing the DM, the process history metadata describing all choices and processes of the activity is generated. The technical metadata and digital provenance (process history) metadata implementation use the Audio Engineering Society's AES-X098B:

(D098B) Audio object structures for preservation and restoration, and AES-X098C: (D098C) Administrative metadata for audio objects – Process history schema, encapsulated in a METS (Metadata Encoding & Transmission Standard) schema. A limitative amount of sound metadata is also included in the Broadcast Wave file (in the Bext-Chunk). Since this metadata is insufficient for sound preservation, it serves for sound data exchange for automation and information exchange between applications.

Given the huge volume of files and metadata generated, an XML plug-in to enable the full automation of the ingest capability according to the established structure was developed. It allows data and metadata to be validated before ingest and stored in a specific level in the hierarchy. Search results are thus more efficient and can be returned at a very precise level. This helps to optimize work processes and increases flexibility, productivity and efficiency.

Digitization generates a huge amount of data that must not only be saved and archived, but must also be available for reuse in order to improve the accessibility of our collection. This solution also manages archiving media based on their use and life cycle, and will allow orderly and possibly automated data migration.

5. STRATEGIES

To be successful, the NFB's mass digitization project requires innovation, a review of current procedures, new work methods and possibilities, and appropriate processing choices based on the content and source medium of each work.

Process automation is one of the keys to success. Although some processes and choices require, and will continue to require, a technician's involvement and manual operation, a large portion of this work will be automated. Automation helps to optimize work processes through formalization and standardization. Automation is implemented progressively with the digitization plan and improves efficiency, particularly in information sharing, data integrity and processing throughput. New ways and methods are explored and implemented in the various workflows, especially for automating colour grading, sound assembly working from sound-signature recognition, image and sound restoration, and certain aspects of quality control.

For example, the implementation of quality assurance processes in various phases of workflows ensures planned and systematic controls for verifying and maintaining levels of quality according to objectives, and for validating operations to ensure that all elements are up to our standards and correspond to the metadata. The reports generated by these processes serve as performance indicator measurements.

A quality control workflow involving a complete viewing of the work to confirm and validate all choices is necessary. The integration of digital technology and the implementation of automation will transform our traditional approach in production and distribution and change the way the quality controls interact with the rest of the production chain.

The goal of the various ongoing automation developments is to organize a series of methods and technics to achieve efficiencies, to have employees work on creative initiatives, and to bring the finalized DM work directly to the quality control level. Only at this point will a human resource validate the automation and, if necessary, bring back the work to manual correction using a more traditional approach. This approach will not be entirely feasible for all works, but it will provide a more efficient way of processing the huge amount of data.

To bring these strategies to fruition, the NFB will implement a number of innovative workflows. The purpose of the workflows is to process the largest number of works possible. Specific workflows will be established to handle exceptions and other difficult cases. This approach looks promising and will enable us to restore works in a way that respects the originals.

6. DIGITIZATION WORKFLOWS

The digitization plan workflow model allows us to represent future optimal workflows with current procedures. These workflow models are organic and will serve as a baseline for work, training, assessment, refinement and continual improvement. Over time, processes will be adjusted, optimized and evaluated for their potential to be automated.

We also detail several underlying procedures to these workflows that support collection digitization, processing and accessibility. When developing workflows, the methodology used is:

- Investigate industry standards, analyze their use in similar contexts and select which to use;
- Analyze all information on the selected standards in order to determine relevant metadata and information structure;
- Analyze and understand existing processes and the limits and potential of current technology in order to handle mass digitization of the collection;
- Check digitization and processing tools to ensure they efficiently capture all metadata connected with the choices made;
- Implement work processes that will allow us to preserve the work's component parts in their current state, but also allow us to reprocess them in the future if necessary;
- Define and implement a data structure for each process with the goal of optimizing our knowledge of the digitized content and search capabilities;

- Design applications that automate processing, perform quality checks and ensure operational efficiency while limiting manual errors;
- Analyze and define the links between work processes and their interoperability.

To facilitate interoperability between systems, technologies and partners, and to ensure efficient access to data in the future, we have selected open, non-proprietary standards that are well established and accepted in the industry. We will install ready-to-use, commercially available infrastructure that will be customized to our specific needs.

The digitization plan opens new opportunities in terms of innovative workflows, advanced technologies, infrastructure modernization, resource training, organization and culture change, and process review. All these developments will progress into all areas of production and distribution.

7. IMPLEMENTATION OF DIGITIZATION PLAN

The integration of digital technology that has been underway for several years now is transforming our production and distribution chains and giving us an opportunity to implement a dynamic digitization and accessibility solution. End-to-end integration of a fully digital production chain also gives us more flexibility. With this flexibility, we can automate workflows when appropriate and modify them as technological capabilities and business needs change.

Over the past five years, the NFB has made more than 6,000 titles accessible in a variety of digital and encoding formats. The NFB has always responded to the accessibility needs of its various clients, and its digitization plan is no different in this regard.

8. TECHNOLOGICAL CHANGE IN YEARS TO COME

Digital screening and audiovisual technologies are revolutionizing the film, entertainment and education industries. We are seeing an explosion of new devices that enable consumers to access the content of their choice. For content distributors, the array of formats that must be prepared for each distribution channel is increasing, while the time available to control each version is diminishing. We must therefore seek more efficient methods to meet this demand.

The choice of using a mezzanine file with metadata and files containing assembly information will enable us to efficiently meet all current and future demand. No file specific to a given distribution channel will be kept, since it can be recreated on request from the mezzanine file. We are confident that this strategy will reduce the volume of data archived for each work and provide an efficient method for automatically generating any new deliverable.

9. REFERENCES

Advanced Media Workflow Association, aafassociation.org.

AES-X098B: (D098B), “Audio object structures for preservation and restoration”, aes.org.

AES-X098C: (D098C), “Administrative metadata for audio objects - Process history schema”, aes.org.

ANSI/SMPTE 268M-2003, Digital Picture Exchange (DPX), smpte.org.

Casey M., Gordon B. *Sound Directions - Best Practices for Audio Preservation*, Indiana University and Harvard University, 2007, http://www.dlib.indiana.edu/projects/sounddirections /bestpractices2007/.

Devlin, B., Wilkinson, J., *The MXF Book*. Focal Press, Burlington, MA, 2006.

EBU Recommendation R111-2007, “Multichannel use of the BWF audio file format (MBWF)”, ebu.ch.

EBU – TECH 3306, « MBWF / RF64: An extended File Format for Audio », 2009, ebu.ch.

EBU Tech 3285, « BWF - a format for audio data files in broadcasting”, 2001, ebu.ch.

EBU Tech 3285-s1, “EBU Tech 3285 Supplement 1: BWF - MPEG Audio”, 1997, ebu.ch.

EBU Tech 3285-s2, “EBU Tech 3285 Supplement 2: BWF - Capturing Report”, 2001, ebu.ch.

EBU Tech 3285-s3, “EBU Tech 3285 Supplement 3: BWF - Peak Envelope chunk”, 2001, ebu.ch.

EBU Tech 3285-s4, “EBU Tech 3285 Supplement 4: BWF - link chunk”, 2003, ebu.ch.

EBU Tech 3285-s5, “EBU Tech 3285 Supplement 5: BWF - chunk”, 2003, ebu.ch.

EBU Tech 3285-s6, “EBU Tech 3285 Supplement 6: BWF - Dolby Metadata”, 2009, ebu.ch.

METS - Metadata Encoding & Transmission Standard, www.loc.gov.standards/mets.

SMPTE 277-1-2009, “Material Exchange Format (MXF) – File Format Specification”, smpte.org.

Session 10a: Cost Models

LIFE^3: A PREDICTIVE COSTING TOOL FOR DIGITAL COLLECTIONS

Brian Hole
The British Library
St Pancras, 96 Euston Road, London,
NW1 2DB,
United Kingdom

Li Lin
The British Library
St Pancras, 96 Euston Road, London,
NW1 2DB,
United Kingdom

Patrick McCann
HATII
11 University Gardens
University of Glasgow
Glasgow ,G12 8QH,
Scotland

Paul Wheatley
The British Library
Boston Spa,
Wetherby, West Yorkshire,
LS23 7BQ,
United Kingdom

ABSTRACT

Predicting the costs of long-term digital preservation is a crucial yet complex task for even the largest repositories and institutions. For smaller projects and individual researchers faced with preservation requirements, the problem is even more overwhelming, as they lack the accumulated experience of the former. Yet being able to estimate future preservation costs is vital to answering a range of important questions for each. The LIFE (Life Cycle Information for E-Literature) project, which has just completed its third phase, helps institutions and researchers address these concerns, reducing the financial and preservation risks, and allowing decision makers to assess a range of options in order to achieve effective preservation while operating within financial restraints. The project is a collaboration between University College London (UCL), The British Library and the Humanities Advanced Technology and Information Institute (HATII) at the University of Glasgow. Funding has been supplied in the UK by the Joint Information Systems Committee (JISC) and the Research Information Network (RIN).

1. INTRODUCTION

Life Cycle Collection Management has been described as "a very complex subject with many practical, financial and strategic interdependencies" [18]. The LIFE model and tool make an important contribution to approaching this subject by providing costing estimates for the lifecycle of digital collections, and consequently allowing for the exploration of the practical and strategic dimensions as well. Stakeholders with an interest in this area include libraries, archives and museums, as well as research and Higher Education (HE) institutions along with the individual researchers within them. As part of their mandate to provide access to their collections for the long term, the greatest concerns that they have involve collection management, technology strategy, human resource management, and central to all of these, budgeting and funding.

The following are examples drawn from recent literature of where costing information could be used to address questions in each of these areas. With a continual influx of material, libraries are constantly forced to make difficult decisions regarding the balance of their collections, such as whether to retain less used physical items due to pressure on storage space [6]. Knowing the true cost of digitising items is important when comparing this to other options such as continued physical storage, disposal and reassignment of space for other purposes [13].

In terms of technology strategy, digital repositories are becoming extremely important as central components of institutions' technology infrastructures [7]. Knowing the relative costs is essential in choosing the correct repository and preservation system, where the future financial consequences of mistakes can be serious [16].

Institutions are often unsure as to their human resource requirements as the digital proportions of their collections increase. Should the related work be done in-house, outsourced, in collaboration with other organisations [11], or by re-training existing staff [17]?

Determining the true cost of a digitisation project and being able to justify it is critical, as most institutions have to seek external funding for such work [5]. Not taking medium and long-term preservation factors into consideration can create a "ticking time bomb" [21], which requires additional, unplanned funding to diffuse at a later date. Organisations need to understand that funding for digital preservation needs to be provided on an ongoing rather than temporary basis, and how to incorporate planning for this into their budgets [9]. Grant applications need to include sound design, a detailed management plan (including a commitment to preservation), a complete and realistic budget, details of all required human resources, and plans for effective sustainability [12].

Institutions require an understanding of the costs of the entire digital lifecycle in all of the above situations in order to ensure sustainability and preservation [3],

especially as the preservation actions involved at each stage may not be initially obvious to them [4]. The LIFE model and tool provide an accessible and practical way of determining these costs, in order that these critical decisions can be made with greater confidence.

2. BACKGROUND

The LIFE project has so far run over a total of three and a half years, spread over three phases. The first phase ran from 2005 to 2006. This established that a lifecycle approach to costing digital collections was applicable and useful, and developed a methodology for doing so. It tested this approach by applying it to real life collections in a number of case studies, including Voluntarily Deposited Electronic Publications (VDEP) and web archiving at the British Library, and the e-journals repository at UCL. It also developed a model for estimating the preservation costs of a digital objects lifecycle [10].

This was followed by phase two in 2007 and 2008, which included further validation of the model, economic assessment of the LIFE approach and further testing and evidence generation via additional case studies. These included the SHERPA-LEAP institutional repositories, SHERPA-DP digital preservation services, and the British Library Newspapers digitisation project. Feedback from the LIFE[2] final conference indicated considerable demand for a predictive costing tool to aid in planning digital preservation [1].

3. LIFE[3]

The third phase which ran from 2009 to 2010 and has just completed, has delivered a web-based predictive costing tool that significantly improves the ability of organizations to plan and manage the preservation of digital content. This tool is based upon a refined version of the LIFE model produced in phase two (see figure 1), following collection of additional case study and survey data. This has enabled the model to cover a wider range of preservation scenarios, including sound, web and e-journal archiving, in addition to print.

Creation or Purchase	Acquisition	Ingest	Bit-stream Preserv.	Content Preserv.	Access
Creation	Selection	Quality Assurance	Repository Admin.	Preserv. Watch	Access Provision
	Submission Agreement	Metadata	Storage Provision	Preserv. Planning	Access Control
	IPR & Licensing	Deposit	Refresh	Preserv. Action	User Support
	Ordering & Invoicing	Holdings Update	Backup	Re-ingest	
	Obtaining	Reference Linking	Inspection	Disposal	
	Check-in				

Figure 1. The LIFE model

A survey of digital preservation repositories was carried out in order to better understand their storage requirements and costs, with these being correlated to the size and purpose of each system. Aside from the number of mirror sites employed, the survey looked at the combination of storage technologies used for access as well as backup, the cost and expected lifetime of the hardware, and also as other factors such as support, infrastructure and electricity costs.

3.1. Model Development

This data was then collected and built into a financial model, using Excel and Visual Basic. The Excel workbook includes a basic input sheet, the output sheet which displays the calculated costs for all the stages, six data refinement sheets that allow the user to modify estimations used within each model stage, and six model sheets that contain the financial models used for calculating costs throughout the lifecycle. The Visual Basic code involves a number of subroutines that are linked with macros to perform functions such as filling and clearing input cells within the workbook.

While the model is designed to produce accurate estimates due to a thorough understanding of the preservation lifecycle and associated variables, it was felt that it should also be able to provide quicker estimates for the purpose of comparison, where many options under consideration can be quickly discounted. A template approach was followed to allow the user to select from content and organisation categories into which their particular project falls. The model is then populated with default data calculated from the mean values of case studies that also fall into those categories.

A user thus has to enter data into only five fields on the basic input sheet in order to receive an in initial cost estimate. These are simply the time frame of the project, the original media type of the material to be preserved (print, website, sound, research material, or other) the source (purchased, donated or to be created through digitsation or harvesting), the number of items to be processed in each year of the project, and the size of institution involved. In the case of digitisation, they are also asked for the quality required. This information is used to pre-populate the model with data averaged from relevant case studies where it is available, and the user is immediately presented with a cost estimate on the output page. They are able to drill down and change the default values at each stage of the life cycle in order to achieve a more precise result using the refinement sheets, or they can simply reset the model and try a different configuration (see figure 2). All figures on the output page are rounded to two significant figures in order to underline the fact that they are indicative estimates only, and users are made aware of the fact that case study data is illustrative rather than absolute. Initial numbers are likely to be higher than expected because it is assumed

that all stages of the lifecycle are being carried out, often defaulting to more conservative scenarios.

The first thing the user is likely to do is adjust the model to use the infrastructure and staff costs specific to their institution. The 'Refine Organisational Profile' sheet will contain default data based on the size of institution the user has selected on the input sheet, but unless the sheet has been previously modified by someone in the same organisation these are unlikely to be accurate. Users can choose the number of storage sites to be modelled, along with the storage technology and its cost for each one, as well as for backup. Technologies included in the model are spinning disk, enterprise tape, flash storage, and pay per use (e.g. cloud storage). One of the highest security factors for preservation is diversity of storage methods and vendors [14], so the ability to experiment with different scenarios and supplier costs here is very useful. Staff costs based on annual, daily or hourly rates should also be entered here for the five project roles used in the model, from Senior Manager to Operational Staff. These rates are used throughout the model wherever staff costs are calculated. For UK HE institutions, users can also enter the indirect and estate figures for each role to ensure proper calculation of Full Economic Costs (FEC). Staff costs are then adjusted for inflation across time.

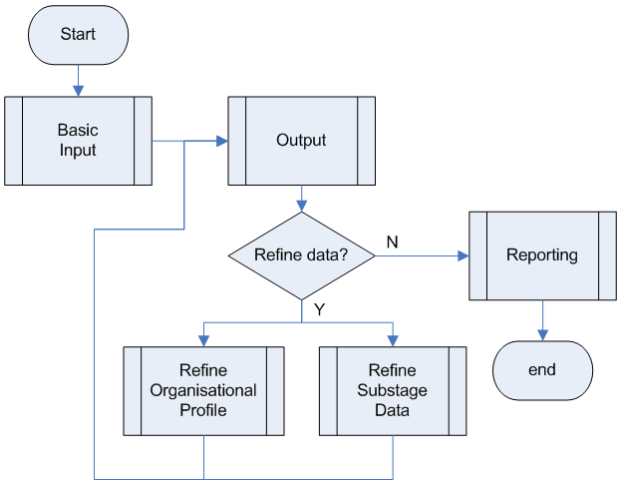


Figure 2. Typical workflow

The 'Creation or Purchase' stage calculates costs based on the source chosen by the user on the input sheet. For purchased items, the total purchase cost is derived by summing-up the purchase cost of all years (purchase cost per item x number of items per year). For donated items, the cost at this stage is simply zero. For digitised items, the user is presented with 23 digitisation cost elements across three columns to capture small, medium and large projects, and the associated case study derived default data. Elements are either based on labour costs (e.g. days of work for a project manager to shape the project) or cost per item digitised (e.g. deshelving and capture). Users should check each one of these figures, correcting them where necessary or setting them to zero when a task is not part of the project under consideration. This challenges institutions to justify non-inclusion of best practice tasks such as QA and metadata capture.

For the 'Acquisition' and 'Ingest' stages, the user is able to adjust the default data for 35 cost elements, based on hours, days, or percentage of time spent on each by staff members of a certain role. As ingest is an area where the KRDS2 project noted that there are potential savings to be made by many projects [2], users should use this section to experiment and try to find cost savings.

The 'Bitstream Preservation' stage allows the user to edit the costs for repository administration, refreshment, backup and administration. In addition to this, the costing factors for each type of storage technology can be changed, including lifetime, cost per MB, rate of cost deflation (applied throughout time) and electricity costs. As the latter cost is especially significant for enterprise systems [14] users should pay attention that this is correct for their region or institution. It is important to note that the technologies we employ today are not permanent solutions however [11], and that we really cannot predict what will be available in 20 years [19], so all model predictions beyond this point should really be accepted with great caution.

It was noted at the end of the LIFE^2 phase that the 'Content Preservation' stage still required development [1], and this has now been simplified and reworked, taking into consideration the work of the Danish national library and archives [8]. Each content type is assigned a heterogeneity level describing the number of different file formats involved of high (e.g. websites) or low (e.g. print), and a complexity level regarding these files of high (e.g. MS Word or PDF documents) or low (e.g. tiff files). The combination of each of these factors is then used to determine the cost of any content migrated. Users are given three migration strategies to choose from, these being 'do nothing', 'migrate on ingest', and 'migrate periodically'. In the case of 'do nothing', users can also enter a cost for emulation. This is the chosen strategy for the KB in Holland for example, betting on the stability of emulation in the long term [14]. The Welcome Library on the other hand count on the fact that by accepting only a limited range of formats thought to be stable, the 'do nothing' option will work without emulation [20]. For 'migrate on ingest', the cost of migration is calculated for each year of the project based on the number of items selected. In the case of 'migrate periodically' the user can determine the percentage of items to be migrated and the number of years between migrations. It is recommended that users challenge their assumptions and experiment with these options, as the costs within this section of the model can be significant depending on the options chosen. It has been noted that institutions should not count on the falling cost of

storage, as a growing number of items due to migration can easily offset these gains [4], while the operational costs of some preservation strategies may actually exceed the perceived value of a collection [14]. Rusbridge has also cautioned that the assumption that file formats become obsolete rapidly and that interventions should thus be made on a frequent basis is likely to be incorrect in many cases where until recently it was an accepted truth [15].

Finally, the 'Access' stage provides default estimates for the costs of creating, maintaining and managing an access system, based on both direct costs and staff effort. Users are also able to determine whether some costs will recur periodically due to replacement or refreshing of the system.

The LIFE3 model has been exposed to members of the digital preservation community during its development, and has received very positive feedback, in particular due to its immediate usability.

3.2. Web Tool Development

In conjunction with HATII, a web-based tool incorporating the financial model has been produced. The aim of the tool is to make the LIFE model both easily accessible and easy to operate for all levels and backgrounds of users. As an example of this, when using the tool in comparison to the spreadsheet, only the data that is directly relevant to the user at any point in time is displayed. Once the user has drilled down into the data and edited it to the point that they feel it is representative of their project, they are able to produce a full report of the predicted cost and all of the factors that have been involved in calculating it. This can not only be used to demonstrate the thoroughness of the prediction, but is a useful checklist for users to make sure that they have in fact taken all required tasks into account.

The application has been developed using the open-source Symfony (http://symfony-project.org) object-oriented PHP framework on top of a MySQL database. PHP and MySQL are well-established open-source technologies in which the developers at HATII have plenty of experience. The use of an MVC framework allowed the development to proceed more rapidly and provided a standard, well-documented structure.

In order to ensure ease of sustainability for the tool in future, it was required that the economic model employed by the application be able to be edited by an administrator without the need for a developer. This meant that as much of the logic of the model as possible had to be contained within the database, the structure of which was kept as general as possible.

The following can be considered a description of the model in the context of the application in the broadest terms possible. A preservation project takes place over a number of years. It is classified in a number of ways (category, source, organisation type etc.) and a number of items are processed each year. The model can be thought of as a set of properties that can be used to describe a project. The value of a property of a given project can be drawn from a case study, entered by the user or calculated from the values of other properties of the project. The ways the project is classified determines which properties apply and how their values are determined.

The basic entities can be seen in this description: project, project year, property, value, classification and category. But much of the power of the model comes from the way in which property values are derived from each other through calculations. To provide the necessary configurability therefore, those calculations also needed to be stored in the database. Simple arithmetic formulae using the sum, product, difference and quotient operators can be easily evaluated when they are described using postfix notation (http://scriptasylum.com/tutorials/infix_postfix/algorithms/postfix-evaluation/index.htm). The algorithm involves reading the expression from left to right, so the formulae are stored in the form of a linked list of components in the database. Each component is either an operand or an operator, and where it is an operand it contains a reference to the property whose value is to be used in the application. The postfix evaluation algorithm can hence be applied quite simply.

A challenge involved in this approach is that the performance of the application can be adversely affected by the need to retrieve not just data for calculation input but the calculations themselves from the database as they are evaluated. Also, any of the values supplied as operands to a calculation may have to be calculated themselves. Another issue is that many properties need to be assigned values for each year of a project, so the number of entities involved in the calculation of an estimate increases greatly as the length of the project increases. PHP's limitations when it comes to managing memory use when executing object-oriented code (specifically garbage collection of objects containing circular references) means that every opportunity needs to be taken to avoid creating objects in memory and to destroy them correctly once they are finished with.

Some specific aspects of the model have had to be handled differently to the standard calculation structure described above. The application of economic factors to costs that recur over each year of a project and costs that occur on a periodic basis are two examples. This logic has therefore had to be written into the application, though the recurrence period and the economic factors themselves remain customisable.

Generally, however, the approach to complications not catered for by the implementation of the model has been to increase the flexibility of the model rather than to implement specific solutions. For example, as it became apparent that additional types of classification were necessary, and that an administrator would need control over them (e.g. organisation size was been added

to the model after development began), these were abstracted away from the project object, allowing the behaviour of the model to be tailored according to all of the possible combinations of classification applied to the project.

Most importantly, the tool has been designed to be easily maintainable by its hosting institution, without the need for further programming. All variables and formulas used in the model can be edited through a user administration interface. In this way the financial model can be modified to take account of new factors (for example a new task or additional hardware requirement) and any errors in formulas can be fixed.

4. FUTURE WORK

While LIFE has to date produced an extremely valuable resource for the digital preservation community, future work will ensure that this resource is widely available and of maximum use going forward. This will focus on making the LIFE tool widely available as a working and sustainable service with promotion, support and knowledgebase maintenance and enhancement. It will also make the service more applicable to a wider range and type of institutions globally, by internationalizing the financial model and extending the breadth and depth of the data.

To do this, LIFE will partner with the Open Planets Foundation (OPF), a new foundation with a global footprint that is dedicated to providing technology, advice and on-line complimentary services for the planning of digital preservation. OPF will provide hosting, promotion, support and maintenance, effectively taking LIFE from a functioning tool to a working, sustainable Service.

The Service will be further developed based upon controlled evaluation with selected HE/FE partner sites, and the accuracy of LIFE cost estimation will also be enhanced by establishing a process for collating and integrating new costing data in the LIFE knowledgebase. Finally, the LIFE Service will also be internationalized in order to improve its usability worldwide, with support for different currencies and a wider range of international data.

5. REFERENCES

[1] Ayris, P. and Davies, R. and McLeod, R. and Miao, R. and Shenton, H. and Wheatley, P. *The LIFE2 final project report.* http://eprints.ucl.ac.uk/11758/ (accessed 11th July 2010), LIFE Project, London, UK, 2008.

[2] Beagrie, N., Lavoie, B. and Wollard, M., *Keeping Research Data Safe 2*, http://www.jisc.ac.uk/media/documents/publications/reports/2010/keepingresearchdatasafe2.pdf (accessed 11th July 2010), Charles Beagrie Ltd., 2010.

[3] Bradley, K., "Defining Digital Sustainability", *Library Trends*, 56(1):148-163, 2007.

[4] Chapman, S. "Counting the Costs of Digital Preservation: Is Repository Storage Affordable?", *Journal of Digital Information*, https://journals.tdl.org/jodi/article/viewPDFInterstitial/100/99 4(2): 1-15, 2004.

[5] Eden, B., "Getting Started with Library Digitization Projects: Funding Your First (and Subsequent) Digital Projects", *The Bottom Line: Managing Library Finances*, 14(2):53-55, 2001.

[6] Holt, G. E., "Economic Realities in Optimizing Library Materials Access", *The Bottom Line: Managing Library Finances*, 20(1):45-49, 2007.

[7] Jacobs, N., "Institutional Repositories in the UK: The JISC Approach", *Library Trends*, 57(2):124-141, 2008.

[8] Kejser, U. B., Nielsen, A. B. and Thirifays, A., *The Cost of Digital Preservation: Project Report v. 1.0*, Danish National Archives and Royal Library, 2009.

[9] Lavoie, B. and Dempsey, L. "Thirteen Ways of Looking at… Digital Preservation", *D-Lib Magazine*, 10(7/8), http://dlib.org/dlib/july04/lavoie/07lavoie.html (accessed July 5th 2010), 2004.

[10] McLeod, R. and Wheatley, P. and Ayris, P. *Lifecycle information for e-literature: full report from the LIFE project.* http://eprints.ucl.ac.uk/1854/ (accessed 11th July 2010), LIFE Project, London, UK, 2006.

[11] Middleton, K., "Collaborative Digitization Programs: A Multifaceted Approach to Sustainability", *Library Hi Tech*, 23(2):145-150, 2005.

[12] Ray, J., "Digitization Grants and How to Get One: Advice from the Director, Office of Library Services, Institute of Museum and Library Services", *The Bottom Line: Managing Library Finances*, 14(2), 2001.

[13] Robinson, C. K., "Library Space in the Digital Age: The Pressure is on", *The Bottom Line: Managing Library Finances*, 22(1):5-8, 2009.

[14] Rosenthal, D. S. H., Robertson, T., Lipkis, T., Reich, V. and Morabito, S., "Requirements for Digital Preservation Systems: A Bottom-Up Approach", arXiv:cs/0509018v2 [cs.DL], 2005.

[15] Rusbridge, C. "Excuse Me… Some Digital Preservation Fallacies?", *Ariadne* 46, 2006.

[16] Seadle, M., "The Digital Library in 100 Years: Damage Control", *Library Hi Tech*, 26(1):5-10, 2008.

[17] Shenton, H., "From talking to doing: Digital preservation at the British Library", http://dx.doi.org/10.1080/13614530009516807 (accessed July 11th 2010), *New Review of Academic Librarianship*, 6(1): 163-177, 2000.

[18] Shenton, H., "Life Cycle Collection Management", *Liber Quarterly* 13:254-272, 2003.

[19] Steele, K., "The Fiscal Wonders of Technology", *The Bottom Line: Managing Library Finances*, 22(4):123-125, 2009.

[20] Thompson, D., "A Pragmatic Approach to Preferred File Formats for Acquisition", http://www.ariadne.ac.uk/issue63/thompson/ (Accessed July 11th 2010), *Ariadne* 63, 2010.

[21] Wheatley, P. and Hole, B., "LIFE3: Predicting Long Term Digital Preservation Costs", http://www.life.ac.uk/3/docs/ipres2009v24.pdf (accessed 11th July 2010), paper presented at iPres 2009.

BUSINESS MODELS AND COST ESTIMATION: DRYAD REPOSITORY CASE STUDY

Neil Beagrie
Charles Beagrie Ltd
Salisbury, United Kingdom
www.beagrie.com

Lorraine Eakin-Richards
School of Information & Library Science, University of North Carolina
Chapel Hill, NC, USA
www.lorraineeakin.com

Todd Vision
National Evolutionary Synthesis Center
Durham, NC, USA
www.nescent.org

ABSTRACT

Data attrition compromises the ability of scientists to validate and reuse the data that underlie scientific articles. For this reason, many have called to archive data supporting published articles. However, few successful models for the sustainability of disciplinary data archives exist and many of these rely heavily on ephemeral funding sources.

The Dryad project is a consortium of bioscience journals that seeks to establish a data repository to which authors can submit, upon publication, integral data that does not otherwise have a dedicated public archive. This archive is intended to be sustained, in part, through the existing economy of scholarly publishing. In 2009, Dryad commissioned the develop-ment of a cost model and sustainability plan. Here we report the outcome of this work to date.

The sustainability efforts of Dryad are expected to provide a model that may be exported to other disciplines, informing the scale needed for a sustainable "small science" data repository and showing how to accommodate diverse business practices among scholarly publishers, funding agencies and research institutions.

1. INTRODUCTION

Researchers, scientists, and publishers recognize that the framework of scientific journal publication is being reinvented as a result of the ascendancy of online access [8]. Sayeed Choudhury of John Hopkins' Virtual Observatory has even argued that the publication of scientific knowledge requires such radically new infrastructure and modes of presentation that while the joint presentation of journals and scientific data may be considered "a new form of compound publication," some situations exist in which "data releases, even without accompanying articles, might be considered a new form of publication" [4]. A host of first mover preservation-oriented organizations and projects such as LOCKSS, Portico, and the DICE group's Storage Research Broker and iRODS infrastructure have set the stage for this new wave of compound publication by creating increased incentives for collaborative preservation of journal articles or research data and by developing solid techniques for ensuring trustworthy preservation. Building from their findings, a number of new initiatives now focus directly upon building the technical and human infrastructure that will enable interoperability between journal articles and associated research data [8][12].

Nonetheless, many of these initial attempts to develop sustainable infrastructure and techniques for creating "enhanced publications,"[1] have focused upon linking large pre-existing databases of research data to the journal articles with which they are associated, such as the Public Library of Science (PLoS) and the Protein Data Bank (PDB) [8][12]. By contrast, the Dryad project focuses on the long tail of datasets reported in the scientific literature that are too heterogeneous in structure to be managed within the (necessarily finite) number of primary bioscience databases.

Because of the frequently more ephemeral and distributed budgetary situations faced by small team scientific publication efforts, the sustainability concerns have required Dryad to develop strategies suited specifically to such efforts. This includes an especially concentrated focus on engaging journal societies and publishers very early in the repository development to ensure that buy-in occurs and is maintained throughout the development cycle. The strategy also includes a large degree of collaboration with institutional partners and like-minded research projects that allow Dryad to take advantage of the inherent cost savings offered by sharing highly skilled personnel and resources.

[1] Woutersen-Windhouwer, S. And Brandsma, R. "Report on Enhanced Publications State-of-the-Art", *DRIVER, Digital Repository Infrastructure Vision for European Research II*, European Union, 2009, p. 7. An enhanced publication is here defined as "a publication that is enhanced with three categories of information: (1) research data (evidence of the research), (2) extra materials (to illustrate or clarify), or (3) post-publication data (commentaries, ranking)." Later in this article, we refer to "supplementary materials and data" to recognize that comparator organizations may include various different scopes of materials when they refer to "data."

2. THE DRYAD REPOSITORY

Dryad (www.datadryad.org) is an initiative incubated by The National Evolutionary Synthesis Center (NESCent), the University of North Carolina at Chapel Hill Metadata Research Center, and the North Carolina State Digital Libraries, who began a working consortium of bioscience journals to develop and sustain a digital repository for publication-related data [10]. The repository was initially developed to help support the coordinated adoption of a policy by a number of leading ecology and evolution journals in which data archiving would be required of all authors at the time of publication [11]. Deposition of data into Dryad is one way of satisfying this policy, although other mechanisms are also allowed or encouraged depending upon journal policy (e.g., data may be hosted by the publisher, or archived in specialized repositories such as GenBank). Journals are responsible for making authors aware of their data archiving policy at the time an article is submitted and enforcing it at the time of publication. The repository software is based on DSpace, which allows Dryad to leverage a technology platform being used by hundreds of organizations and maintained by a large and active open-source software community.

Dryad's start-up funds have come primarily from a four year US National Science Foundation (NSF) grant awarded in 2008, as well as NESCent, the NSF-funded DataONE initiative, and the US Institute for Museum and Library Services (IMLS). In addition, a new award through the Joint Information Systems Committee (JISC) in the UK funds Oxford University and the British Library as development partners in Dryad.

The NSF grant identified as a key goal the need to establish stakeholder ownership and governance of Dryad, where journals serve as key stakeholders. To meet this goal, Dryad has created the Dryad Consortium Board (DCB), a central governing body that oversees the repository's strategic planning and to whom the repository staff report. One of the major tasks of the DCB is to agree to a sustainability plan and to help implement it. This will ensure that Dryad can honour its long-term commitment to data preservation.

The DCB currently operates under an interim governance structure consisting of one voting representative from each partner journal. The requirements for partnership for the period prior to the launch of the service in January 2012 [6] include the following:

- Formal adoption of the Joint Data Archiving Policy[2], or an equivalent policy requiring submission of data as a condition of publication;
- Commitment to the development of a self-sustainable business model for Dryad; and
- Appointment of a representative to the DCB with full voting authority.

The DCB elects an Executive Committee of five journal representatives who, together with the Project Director, are responsible for routine oversight of the repository. The Executive Committee is required to bring major financial and governance decisions to the full board for consideration [6].

In addition, partner journals are expected to share article metadata with Dryad prior to publication, to provide information to authors on how to submit to Dryad at the time of submission or acceptance, and to include links to Dryad data within the respective published article, as in [9].

3. A COST MODEL FOR DRYAD

Lorraine Eakin-Richards was commissioned in October 2009 to prepare an initial cost model to help estimate expected repository costs in preparation for sustainability discussions at the DCB meeting in December 2009.

The aim of the cost model was to provide the board with a better understanding of the cost components that the Dryad repository could expect to encounter in both its initial stages of operation and during the early stages of growth expected over a five year time frame. It also provided initial estimates of total and per paper costs.

Eakin-Richards worked with the Dryad project team to assess these current and projected costs, to identify potential cost-share elements from likely ongoing line item budget categories, and to provide a worksheet that could be used on an ongoing basis by the project team to fine tune initial estimates. The key cost components and philosophy of the model were derived, after review of numerous previous cost modeling studies, from the JISC *Keeping Research Data Safe* model [1], which appeared most closely to map to the requirements of the Dryad repository cost modelling needs. The structure of the worksheets was based upon the activity-based cost model built by Eakin and Pomerantz [7].

High level cost categories and potential detailed line item breakdown of these categories were made available to help Dryad begin to project likely costs over time and to fine tune initial estimates as the DCB finalizes its strategies and policies. This breakdown can be viewed in Table 1. Some categories were deemed unnecessary for Dryad's particular situation, such as research and development costs for service innovation, which are expected to be covered via grant funding, and infrastructure costs, which are part of Dryad's

[2]The Joint Data Archiving Policy (JDAP) is a policy of required deposition to be adopted in a coordinated fashion by Dryad partner journals [11]. By adopting the JDAP, a journal agrees to require that data used in support of the conclusions of an article be submitted to a suitable public repository as a condition of publication. Some exceptions hold. For example, authors may elect to embargo access to the data for a period of up to one year following publication of an article; exceptions can also be granted at the discretion of the journal editor in situations such as those in which the data may contain sensitive information regarding human subject data or the location of endangered species.

institutional partner cost sharing arrangements. Any repository wishing to emulate these categories should select those line items most relevant for its own environment and purposes.

One recommendation of the cost modelling consultancy, however, was that a full cost assessment be made and that a risk assessment and strategy be developed to cover the possibility that any particular cost share element be reduced or lost due to budgetary emergencies or strategic changes among the partner institutions. In addition, as longer term planning around operational costs occurs, Dryad will benefit from engaging in time discounting of expenses, in order to gain a better understanding of its "true" long-term economic costs [7].

Repository Management
Repository Manager Salary and Benefits
Advisory Board Meeting Costs
Administrative Support
Administrative Support Salary and Benefits
Curation
Lead Curator Salary and Benefits
Curator Salary and Benefits
Storage and Hardware
System Administrator Salary and Benefits
Hardware Refresh
Security Services
Infrastructure/Facilities
Ongoing Space Expenditures
New Furniture and Equipment Expenditures
Network Set-Up and Maintenance
Telephone and Communications
Research and Development
Personnel Salary and Benefits
Personnel Travel (Specifically Related to Research Collaboration)
Repository Cost Share on Collaborative Projects
Repository Maintenance
Developer Salary and Benefits
Technical Manager Salary & Benefits
Software Expenses
Outreach and Promotion
Communications Specialist Salary and Benefits
Travel for Communications Purposes (e.g., Vendor negotiations, conducting training & workshops)
Advertising Charges
User Documentation and Training
Personnel Salary and Benefits
Outsourcing
Vendor and Consulting Fees
Miscellaneous
Personnel Travel
Personnel Training
Communications Costs (Management, Outreach, Advisory Board, Telephone call charges, etc.)
Miscellaneous Supplies
Insurance
Contingency Estimate

Table 1. Potential Cost Components

The overall projected costs of the service will vary according to the level of investment in value-added services (i.e., data curation). With low to moderate curation effort, initial projections of potential costs for Dryad lead to ballpark estimates of $200,000 or $320,000, respectively, assuming receipt of data from 5,000 or 10,000 papers per annum.

3.1. "Per Paper Costs"

Per paper costs were included within the cost model in order to aid the DCB in determining the feasibility of potential cost recovery techniques. Because buy-in and financial cost sharing from partner journals is a key component of the sustainability model, the journals, societies and publishers needed detailed information about the projected costs, and the repository needed information about what scale of service would be financially viable. The DCB also deemed that a fair model of cost-recovery from journals would need to account for both per paper costs and for the variable number of papers published by each journal in a given year. Given the budget estimates for volumes of 5,000 and 10,000 papers per year, Dryad's per paper expenses were estimated to be $40 and $32, respectively.

3.2. Testing Cost Projections

It is very early in Dryad's development to accurately populate an activity model that could be used to derive full costs for its future activities. In particular, costs for curation will vary according to the level of additional work, e.g., metadata enhancement, and the packaging and documentation for re-use in teaching that may be undertaken by Dryad. Dryad is thus working on the development of a set of "curation service levels" and their associated costs. This is similar to the practice of some publishers, such as the Journal of the American Medical Association or data archives such as the UK Data Archive. Dryad also reviewed use of students or outsourcing to foreign labour markets as part of Dryad's future curation staffing.

4. SUSTAINABILITY PLANNING

In addition to the cost modelling project, Charles Beagrie Ltd was commissioned to work with the Dryad project team to develop sustainability and business planning for the repository. This work began in October 2009 and was completed in April 2010.

The aim was to set a framework in place for future sustainability. Charles Beagrie Ltd incorporated results from Eakin-Richards' cost model and projections by the

Dryad project team, led by Todd Vision, within the framework.

The framework is intended to be a dynamic document that can be maintained, reviewed at least annually, and maybe more frequently over the first 2 years, and will evolve over the life of the project and beyond. It provides guidance on sustainability with the aim of informing business planning. It consists of the following components:

- Strategy, performance indicators and measures
- Comparators and understanding of the costs
- Advantages, benefits and revenue options
- A proposal for sustainability
- Revenue scenarios for Dryad
- Risks register

4.1. External Comparators

We found through desk research and interviews with journals, publishers and data centers that little is known by journals about the specific costs of handling supplementary materials and data. Costs, principally staff time, were observed to vary according to the tasks undertaken and the level of investment in value added services [2]. It is currently not possible directly to compare costs incurred by journals for supplementary data with those for Dryad as they are either largely unknown for the journals, or in Dryad's case, for tasks such as adding metadata to supplementary files, etc., which some journals currently do not undertake. However, it could be observed that the proposed per paper expenses for Dryad appear very reasonable compared to existing author charges (where these exist) for publishing supplementary data files. Amongst three of the journals we interviewed, these author charges for supplementary data files ranged from $100 to $300+.

We also found that while there is no exact archive or repository comparator for Dryad, other archive repositories do offer enough similarities to be of use in comparing some overall costs. Initial analysis, with feedback from the Dryad management team, indicates that a staff of 2-4 FTEs would be a viable initial base level of staffing to deliver Dryad's basic operations. This is comparable to the minimum staffing of other archive comparators at launch we have considered.

The comparators we have reviewed are embedded within larger institutions and can thus leverage pre-existing infrastructure and effort, expertise and direction from associated staff, often co-located but funded separately while working on related activity, including support services, project based research and software development. This helps to maintain a dynamic and sustainable organisation that can respond to change and deal with fluctuations in staffing. Dryad currently is similarly embedded within a larger institution. We noted that Dryad needs to determine the most cost effective way of providing its administration and infrastructure support going forward and ascertain whether support from a host institution can be negotiated at a mutually agreeable cost or provided as an "in-kind" contribution. In due course, a separate not-for-profit legal entity may be considered.

4.2. Transitioning from Project to Service

The transition from Dryad's development phase to a sustainable repository service requires careful planning and the development of a transition strategy. The main considerations revolve around organization and governance; staffing levels; maturity and reliability of automated processes to sustain the repository; and the level of active outreach, training, and member participation to build a critical mass of data available through Dryad. During the transition period, the Dryad team must effectively accommodate changing functional requirements, challenges of scaling the service, and changes in governance. Presently, quarterly repository development plans are reviewed by the Executive Committee, and priorities each quarter are set with careful attention to the needs of current and potential partner journals.

The views of funders on future or continued grant support for Dryad will need to be investigated further. Three categories of grant funding could be important: the possibility of tapered "transition funding" to facilitate the transition from project to service and to allow for the growth of the service in its early years; internationalisation of the service (e.g., mirroring or nodes in Europe or elsewhere) to provide opportunities for widening participation and funding of the service; and research and development opportunities to innovate and enhance the service provided.

Currently 16 interim partner journals participate in the Dryad Consortium. The DCB will consider and agree upon the potential future growth or optimum size of the consortium, appropriate timescales for reaching this size, and impacts on revenues/costs as part of the transition strategy.

5. PROPOSAL FOR SUSTAINABILITY

National or subject repositories are funded in the main through a mixed economy of core and project funding where a maximum of 50% core funding is the norm. Although this can lead to tensions in balancing priorities, diverse revenue streams offer a realistic path for sustaining continued funding and provide some flexibility to decisions around future development.

The easiest and often most successful approaches for projects looking at sustainability issues and possible revenues are to identify those stakeholders that will most benefit from the service and assess their ability and willingness to provide continuing support. Multiple revenue streams can be hard to manage and will bring an additional overhead to the organization that should not be

underestimated, so a necessary balance has to be found between the risk of being dependent upon just one or two revenue streams or that of spreading risk across many but then having to deal with managing them.

At the heart of any sustainability model should be a clear articulation of the "value proposition"– how the organization provides a solution to a problem or delivers an attractive product to its stakeholders and users – that would be otherwise difficult, expensive or impossible for them to obtain [3].

For Dryad the value proposition is as follows:

- For scientists, Dryad will increase citations and the impact of their work. It preserves and makes available data that can be used for more complete meta-analysis, for verification of previous results, and to address novel questions with existing data. Dryad provides an easy mechanism for maintaining data over the long term, thereby facilitating compliance with funding agency mandates;
- For publishers, Dryad frees journals from the responsibility and costs of publishing and maintaining supplemental data in perpetuity, and allows publishers to increase the benefits of their journals to the societies and the scientists they support;
- For funding organizations, Dryad provides an extremely cost-effective mechanism for enabling new science and making funded research results openly available.

For future sustainability, the key questions Dryad faces are: what value can be placed on these solutions and products; what are the size and composition of the communities that will receive benefits; and what should be the size of the Dryad Consortium and of the economies of scale delivered to its participants. The feasibility of sustainability for Dryad will depend upon the following factors:

- The costs of maintaining the Dryad organization and its supporting technology;
- The number of partner journals and the rate at which new data packages are ingested;
- Dryad's success in addressing the varying interests of multiple stakeholders, including journals, scientific societies and publishers;
- The extent to which Dryad increases its visibility in the research community, to which there is increase in the practice of data reuse, and to which there is increased adoption of data citations; and,
- The extent to which Dryad can attract funding / revenue for both operating costs and continued development of its service.

Cost and revenue models together with projections and options that should achieve sustainable services over time were presented in a confidential client report. Key components for maintaining the models and sustainability are ongoing review by the Dryad Consortium Board, regular updates to cost and revenue data, and monitoring and updating of the risks register.

6. THE FUTURE

Upon review of the recommendations from the consultancies, the DCB executive committee drafted a prospectus that is currently being circulated among journals, publishers, scientific societies and funding agencies for feedback [5]. The proposal, which will be reviewed by the full DCB in fall 2010, would establish Dryad as a subscription service by the beginning of 2012. As a major outcome of this work, Dryad is actively expanding the scope of its disciplinary coverage and its institutional partnerships, particularly outside the United States.

7. CONCLUSION

Dryad's funding and development has come at a time when both the sustainability of preservation-oriented programs and the advancement of scientific data repositories has captured the interest of scientists, information science professionals, scientific journals and funding agencies. Dryad's primary aim is to facilitate data discovery and reuse by the research community by ensuring the long-term preservation of the data underlying peer-reviewed articles in the biosciences.

Dryad's business planning efforts are of value beyond the journals and societies directly involved. By testing the idea that both the socio-cultural and economic barriers to data archiving can be overcome within the economy of scholarly communication, Dryad provides a model that may be exported to other disciplines. In particular, it will inform the scope for a sustainable repository, one that balances the need for an economy of scale with the need for cohesion within scientific standards and practices. The model also informs how to accommodate diverse business practices among scholarly publishers, the resources of funding agencies and the capacity of research institutions. To the degree Dryad succeeds in establishing a widely-used and sustainable archive, it will serve as an exemplar for how to realize the full value of the enormous investments in primary scientific data collection.

8. ACKNOWLEGEMENTS

Business planning for Dryad is supported by National Science Foundation grant #0743720. We would like to acknowledge the input of our colleagues Julia Chruszcz, Peggy Schaeffer, and Peter Williams who contributed to the sustainability planning. We would also like to thank the anonymous reviewers of this paper for their most helpful comments and suggestions.

9. REFERENCES

[1] Beagrie, N., Chruszcz, J., & Lavoie, B. *Keeping research data safe: A cost model and guidance for UK universities*, JISC, 2008.

[2] Beagrie, N., Lavoie, B., Woollard, M. *Keeping Research Data Safe 2*, JISC, 2010. Results of the Dryad interviews with publishers will appear in Beagrie, N., Vision, T.J., & Williams, P. *Learned Publishing*, forthcoming.

[3] Blue Ribbon Task Force on Sustainable Digital Preservation and Access. *Sustainable Economics for a Digital Planet: Ensuring Long-Term Access to Digital Information*, San Diego Supercomputer Center, San Diego, 2010.

[4] Choudhury, G. Sayeed. "The Virtual Observatory Meets the Library," *The Journal of Electronic Publishing*, 11(1). http://dx.doi.org/10.3998/3336451.0011.111.

[5] Dryad Consortium Board Executive Committee. *"Subscription plans"* (proposed). https://www.nescent.org/wg/Dryad/images/8/80/Subscriptions20100422.pdf.

[6] Dryad. *Draft Interim "Governance Plan"*, December, 2009. https://www.nescent.org/wg/Dryad/images/1/1d/Governance_for_dec_meeting.pdf.

[7] Eakin, L. and Pomerantz, J. "Virtual Reference, Real Money: Modeling Costs in Virtual Reference Services", *portal: Libraries and the Academy* 9(1):133-164, 2009. DOI: 10.1353/pla.0.0035.

[8] Fink, J. and Bourne, P. "Reinventing Scholarly Communication for the Electronic Age", *CTWATCH Quarterly* 3(3):26-31. http://www.ctwatch.org/quarterly/articles/2007/08/reinventing-scholarly-communication-for-the-electronic-age/.

[9] Lowry R., Urban, E., and Pissierssens, P., A New Approach to Data Publication in Ocean Sciences, EOS, Transactions American Geophysical Union, 90(50):484-486, 2009. DOI:10.1029/2009EO500004

[10] Vision, T.J. "Open data and social contract of scientific publishing", *Bioscience* 60(5):330-331, 2010. DOI:10.1525/bio.2010.60.5.2.

[11] Whitlock, M.C., McPeek M.A., Rausher M.D., Rieseberg L., and Moore A.J. Data Archiving. American Naturalist 175(2):145-146, 2010. DOI:10.1086/650340.

[12] Woutersen-Windhouwer, S. And Brandsma, R. "Report on Enhanced Publications State-of-the-Art", *DRIVER, Digital Repository Infrastructure Vision for European Research II*, European Union, 2009. http://www.driver-repository.eu/component/option,com_jdownloads/Itemid,83/task,view.download/cid,53/.

Session 10b: Strategies and Experiences

SEVEN STEPS FOR RELIABLE EMULATION STRATEGIES SOLVED PROBLEMS AND OPEN ISSUES

Dirk von Suchodoletz
Klaus Rechert
University of Freiburg
Institute of Computer Science

Jasper Schröder
IBM Netherlands
Amsterdam

Jeffrey van der Hoeven
Koninklijke Bibliotheek
The Hague

ABSTRACT

After four years of research within the PLANETS project and two years of KEEP the jigsaw puzzle of emulation becomes a more complete picture. Emulation strategies are now seen as a viable complement to migration. A conceptual and theoretical groundwork has already been laid out, e.g. proper definition and selection of suitable emulators. However, integration into preservation frameworks and additional software archiving remain open research questions. This paper discusses several aspects of reliable integration and proposes development steps for a more complete emulation-based strategies in long-term preservation.

INTRODUCTION

For more than fifteen years there has been a vital debate on using emulation as a strategy to ensure long-term access to digital records. Although emulation has always been an essential addition for many types of digital objects, emulation strategies still have little relevance in practice despite many shortcomings, such as improper handling of dynamic artifacts and authenticity problems in various migration strategies. In contrast to migration, emulation does not require changes to the object or its structure. Hence, the original state of the digital artifact and its authenticity is preserved. However, emulation strategies are considered too expensive and too complex to be a viable solution to address digital preservation challenges [1].

Research on emulation as a long-term archiving strategy matured since the first reports on archiving of digital information in 1996 [7], fundamental experiments with emulation executed by Rothenberg [10] and the theoretical and practical work within the longterm preservation studies of IBM and the Netherlands National Library [13]. The Keeping Emulation Environments Portable project[1] aims to develop a strategy that ensures permanent access to multimedia content, such as computer applications and console games. The main research focus is on media transfer, emulation and portability of software. The platform will allow an organization to capture data from old physical carriers and render it accessible to users by using emulation.[2] To avoid the platform itself from becoming obsolete, a virtual layer guarantees portability to any computer environment [4].

Up to now there has been a strong focus on different emulation concepts as well as strategies to preserve the emulators themselves [14].[3] Especially if looking at the more general emulation approaches, the question of additional software components needs to be taken into consideration (Fig. 1). Additionally, some relevant factors like the integration into emulation frameworks and the cost-effective application of emulation were ignored.

This paper gives an overview on the current status of research and describes requirements and challenges for successful emulation strategies. Therefore, we present a number of solutions like automation of emulation sessions, migration-through-emulation workflows and suggestions of preservation framework integration.

1. STEP 1: SOFTWARE INSTEAD OF HARDWARE MUSEUMS

An obsolete hardware collection is not a viable solution to preserving old computer architectures. The only reason for keeping hardware is to enable access to deprecated media for digital archeology or to present old platforms in a specific setting like a technical or computer games museum. The number of items to

[1]KEEP, *http://www.keep-project.eu*

[2]Requirements and design documents for services and architecture of emulation framework *http://www.keep-project.eu/ezpub2/index.php?/eng/content/download/7918/39623/file/-KEEP_WP2_D2.2_complete.pdf* Specification document for all layers of general-purpose virtual processors, *http://www.keep-project.eu/ezpub2/index.php?/eng/content/download/7917/39619/file/KEEP_WP4_D4.1.pdf*.

[3]Emulation Expert Meeting 2006 in The Hague, *http://www.kb.nl/hrd/dd/dd_projecten/projecten_emulatie-eem-en.html*.

preserve risks becoming too large as it is a necessity to preserve environments for various types of digital artifacts. In addition, the space needed for a larger number of devices together with the energy required to educate and employ a very specialized maintenance crew for every different system

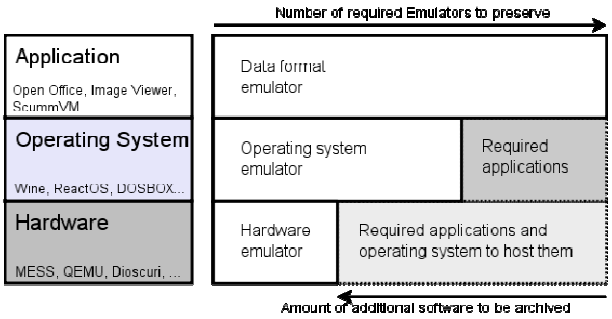


Figure 1. Number of emulators and amount of additional software required depending on the layer chosen

would be a large feat [3].[4] Unfortunately, electronic circuits will not run forever. They will fail at some point independent of the usage pattern. Furthermore, the probability of finding a spare part becomes slimmer each year a platform is out of production. Finally, the concept is fully dependent on the location, meaning there is no easy way to share resources between different memory institutions and users might have to travel large distances for access to a certain system.

Emulation uses a different approach compared to other well-established migration strategies in digital preservation. Emulation strategies usually do not operate on the object itself, but are intended to preserve the object's original digital environment. Emulation helps in becoming independent of future technological developments and avoids the modification of certain digital artifacts in a digital long-term archive.

The concept of emulation is not new to computer science. Emulators have existed for quite some time. Therefore, the list of the developed emulators is astoundingly long and covers a fairly wide range of areas. Prominent examples in the Open Source community are projects like ScummVM (Fig. 2), QEMU, Mess or Mame, just to mention a few. Not every emulator, however, is suitable for the needs of a long-term digital archive. Requirements of the respective archiving organization need to be differentiated, e.g. a national archive requires different computer platforms than a computer games museum. Emulators preserve or alternatively replicate old digital environments in software. They bridge outdated technologies with modern computer environments. Generally, for current computer platforms, three levels for the implementation of emulators can be identified: Topmost the application layer, followed by the operating system layer and on lowest level the hardware layer (Fig. 1). The latter uses the broadest approach, meaning no application and operating system needs to be rewritten to be able to access thousands of different digital object types. The function set of a hardware platform is straight-forward and often much smaller compared to operating systems or applications. Another advantage results from the

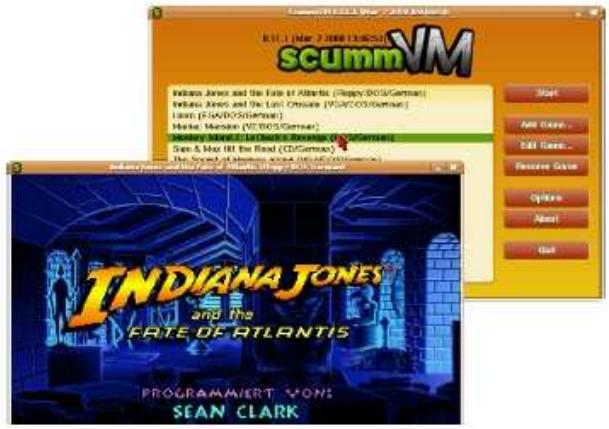


Figure 2. ScummVM is a popular application level emulator for the Lucas Arts and similar type of games

smaller number of hardware platforms in comparison to operating systems.

2. STEP 2: PRESERVING THE EMULATOR

Emulators face the same problems as do every software package and general digital objects. For this reason the considerations of perpetuation of the emulator for future use is a central component of a reliable preservation strategy [15]. Hence, emulators need to be adapted to the current hardware and operating system combinations regularly. The possibility of software migration is achieved through a suitable choice of emulators. If the emulator is available as an Open Source package, it can be ensured that a timely adaption to the respective new computer platform appears. Common and portable programming languages such as C should allow a translation with the respective current compiler. The main advantage of this approach is the use of only one emulation layer.

If there is no possibility to port the emulator on to a new host platform, the recently outdated host platform for which the emulator was created can be emulated [14]. This is referred to as nested emulation. This is a considerable advantage to avoid the complexity of a migration approach.

In the field of hardware emulation and virtualization (e.g. x86 architecture), successful commercial as well as Open Source solutions co-exist. During PLANETS we observed that commercial solutions, like VMware, Parallels of VirtualPC are not suitable for long-term horizons, since the vendors merely follow short-term

[4]Computer Museum Universiteit van Amsterdam, *http://www.science.uva.nl/museum*

interests, e.g. support current operation systems and software. With QEMU and Dioscuri two valid open alternatives exist.

QEMU is a multi platform Open Source emulator implementing x86, ARM, Sparc, PPC and further architectures. It supports a wide range of peripheral hardware from every era of the different platforms. We closely observed the development and advancements of the project throughout the duration of PLANETS and recorded significant advancements. Nevertheless, a number of problems like volatile support of major operating systems occurred and needs to be taken into consideration.

Dioscuri None of the mentioned emulators have been developed primarily for the purpose of long-term archiving of digital objects. This has changed with current research. Dioscuri [11] is a modular emulator which supports both recreation of an x86 computer environment and a durable architecture. With such a design Dioscuri is capable of running on many computer systems without any changes to the software itself. That way, there are chances that the emulator will sustain. At current state Dioscuri can render all kinds of applications from the MS-DOS era. The emulator is developed in Java which runs on top of the Java Virtual Machine and thus is portable to any computer platform that has a JVM running. The internal structure of Dioscuri is very similar to that of common hardware. Each functional entity (e.g. CPU, memory, storage, graphics) is implemented as a software module. Configuring these modules creates a virtual computer.

UVC In the course of research of the last few years we investigated alternate approaches like Universal Virtual Computer [6, 12]. UVC is different as it specifies a computer which is generally available.[5] The intention is to keep the specification of UVC stable over a long period of time. The instruction set of the UVC is limited and during the PLANETS project two new implementation strategies have been developed, one in C++ and one in C. The development effort for each version consisted of roughly four months of work. The expectation is that in the future a new implementation of the UVC can be made in a reasonable amount of time. Having a stable virtual computer layer, preserving the operating system and the applications on top of that ”hardware layer” (Fig. 1) will not be a big issue. These layers will keep doing their jobs. The complexity of the UVC-preserved applications has increased. In the beginning just image conversion applications were available on the UVC. During PLANETS, the logic of the SharpTools spreadsheet was ported as an example for more complex logic. As emulation is no longer a standalone activity and the UVC was made available as a Web service.

3. STEP 3: VIEW PATHS

Digital objects cannot be used by themselves, but require a suitable context to the already mentioned working environment in order to be accessed. This context, called working or utilization environment, must combine suitable hardware and software components so that its creation environment or a suitable equivalent

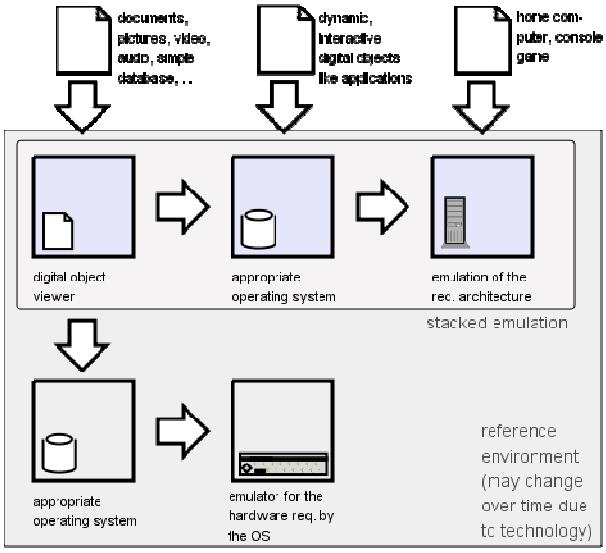


Figure 3. Different view path depending on object typeand archiving strategy

is generated, depending on the type of the primary object. No matter which emulator is chosen, contextual information of the original environment of the digital artefact was created in is always required. For example, questions such as ”for which operating systems is Ami Pro 3.0 compatible with? ” are less obvious today then twenty years ago. To overcome this gap of missing knowledge, a formalization process is needed to compute the actual needs for an authentic rendering environment. In 2002 the concept of a view path [6] was proposed which we refined during research on emulation in PLANETS [15, 16].

A view path reproduces old computer environments or corresponding equivalents as ways from the object of interest to the working environment of the archive user. In other words, a view path is a virtual line of action starting from the file format of a digital object and linking this information to a description of required software and hardware (Fig. 3). Depending on the type of object, a specific rendering application may be required. This application requires a certain operating system to be executed, whereas in turn, it relies on particular hardware.

[5]Alphaworks, *http://www.alphaworks.ibm.com/tech/uvc*

4. STEP 4: ENABLING ACCESS TO EMULATION

In order to allow non-technical individuals to access deprecated user environments, the tasks of setting up and configuring an emulator, injecting and retrieving digital objects in and from the emulated environment have to be provided as easy-to-use services. Making these services web-based allows for a large and virtually global user base to access and work with emulated systems.

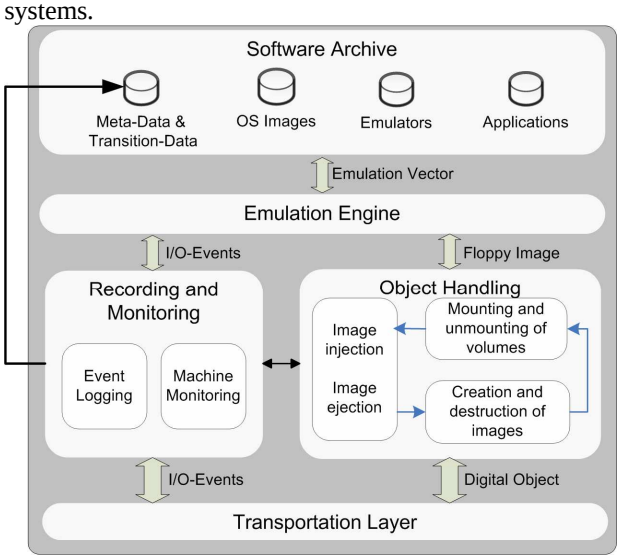


Figure 4. GRATE Architecture

During the PLANETS project we developed the prototype GRATE[6] which allows the wrapping of various software environments within a single networked application. Designed as a general purpose remote access system to emulation services the architecture provides an abstract interface independent of the digital object's type to users and thus was linked to other Web services like PLATO [2].

Screen output and input via mouse or keyboard – which until now are still the most used methods of human-computer interaction – are handled using an event and transportation layer. Currently events and screen output are transferred by using the open and widely used VNC protocol [9]. Figure 4 shows the general architecture of GRATE and its main building blocks. The access to digital objects does not depend on local reference workstations like in archives and libraries. By separating the emulation part from the archive user's environment, GRATE avoids a number of problems, like a sophisticated local installation of a range of software components in unpredictable user environments of different origins. The user does not need to be a trained specialist of ancient computer platforms, but in contrast it is equipped with an user interface similar to many Web 2.0 applications. Furthermore, this approach does not need to transfer proprietary software packages to end- user systems and thus might avoid licensing and digital rights management issues. The management of such services could be centralized and several institutions could share the workload or specialize on certain environments and share their expertise with others. Institutions like computer museums could profit as well, because they are able to present their collections in non-traditional ways rather than simply within their own room, consequently attracting more attention.

Another challenge arises from the transport of the requested artifact from the current into its original environment. Loading of digital objects is a major part of any automated processing setup. The file sets need to be passed into the emulated environment [15]. This is typically a non-trivial task and depends on the feature-set of the original environment. There are two challenges to be faced:

- Network transport from the user's site to the emulation Web service
- Local transport to the target environment

Emulators usually use special container files as virtual disk images. Therefore, they offer an option to transport a digital object into the emulated environment by embedding it in the container file, or by creating a secondary one, which is then attached as an additional virtual hard-disk. However, for producing or modifying such containers, exact knowledge of the internal format is required and usually additional tools are necessary. Furthermore, modifying container files usually cannot be done while the emulator is running, since changes to its internal structure might lead to a corrupt container file. In contrast, floppy and optical disks like CD or DVD are typically removable and thus offer a data exchange option while the emulator is running. Some emulators like QEMU support virtual media loading and ejecting functionality and media changes are noticed by the operating system. Not all hardware platforms and operating systems support optical drives, but most of them support floppy disks.

5. STEP 5: DEALING WITH INTERACTIVITY

A further central problem next to the Framework integration lies in the automation of the human-computer-interaction. Typical digital objects were created with interactive applications on computer architectures with graphical user interfaces. The user was required to point and click using a pointer device (e.g. computer mouse) or using the keyboard to create or modify an object.

The traditional approach supporting the user to automate interactive tasks is the use of so-called macro-recorders. These are specialized tools to capture sequences of executed actions. However, this

[6]GRATE – Global Remote Access To Emulation, *http://planets.ruf.uni-freiburg.de*

functionality is not standardized in terms of its usability and features. Not only are special software components needed, but knowledge on using such applications and operating systems is also necessary.

For a generic approach, a technical and organizational separation between the machine used for executing workflows and its input/output is required. Hence, emulated or virtualized environments are particularly well suited for recording an interactive workflow, such as installing a specific printer driver for PDF output, loading an old Word Perfect document in its original environment and converting it by printing into a PDF file. Such a recording can serve as the base for a deeper analysis and the generation of a machine script for the future than completely automated repetition. By using the aforementioned method, the authors demonstrated the feasibility of such simple migration task in an automated way [8].

An interactive workflow can be described as an ordered list of interactive events. Interactions might be mouse movements or keystrokes passed on to the emulated environment through a defined interface at a particular time. By using a generic approach to describe interactive events, there is usually no explicit feedback on executed interactive events. While a traditional macro-recorder has good knowledge about its runtime environment (e.g. is able to communicate with the operating system), in a generic emulation setup usually the screen output and the internal state of the emulated hardware are the only things visible (e.g. CPU state, memory). Furthermore, the recording/playback system has no knowledge of the system it operates. Hence a framework replaying a complete workflow in a reliable way is indispensable.

A solution relying solely on the time elapsed between recorded actions is not sufficient because executing recorded actions will take different amounts of time to complete depending on the load of the host-machine and the state of the runtime environment. Therefore, we link each interaction with a precondition and an expected outcome which can be observed as a state of the emulated environment. Until this effect is observed, the current event execution has not been completed successfully and the next event cannot be processed. While in the case of human operation the effect is observed through visual control in an automated run, an abstract definition of expected states and their reliable verification is necessary.

One suggested solution makes use of visual synchronization points [17]. For example, a snapshot of a small area around the mouse cursor can be captured before and after a mouse event and then used for comparison at replay time. Hence, replaying an interactive workflow becomes independent of computation time and the host-machine needs to complete a particular action execution. However, removing time constraints still does can not guarantee a reliable playback in general. First, if the synchronization snapshot is done in an automated way, important aspects of the observable feedback on executed actions might get lost. An optional manual selection of the snapshot area recording can improve the reliability since the user is carrying out the recording and is usually familiar with the interaction model of the graphical environment he operates. Second, mouse and keyboard events are passed on to the runtime environment through an abstract interface (e.g. through hardware emulation of a PS/2 mouse interface). Hence, sometimes the environment does not react to input events in the expected way. This occurs for example if the operating system is busy and unable to process input events. For reliable playback, such failures need to be detected and handled by the

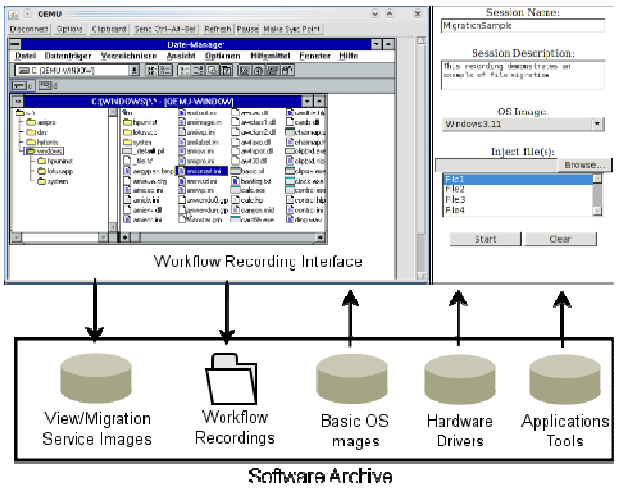


Figure 5. Planets' Web frontend to emulation services needs to be extended with interfaces to software archive

framework. Furthermore, the operator needs support to implement specific failure recovery strategies, e.g. resetting the machines to a stable previous state and retry the failed subsequence. Additionally, if the operator is able to attach meta data to specific events describing its original intend and possible side effects, not only will the reliability of automated execution be improved, but also specific knowledge on practical operation will be preserved.

To support these ideas, the interactive workflow has to be represented as a set of time-independent event transitions, relying only on valid and stable pre- and postconditions. For describing pre- and postconditions, the aforementioned visual snapshot technique was used, but extended to support users choosing the relevant snapshot area. The framework accepts three types of input events: keyboard entry, mouse events and special pseudo-events. Pseudo-events include specific control commands of the runtime environment (e.g. ctrl-alt-del) but might also be used to map the progress of longer running tasks (i.e. installation procedures) through empty dummy events. Mouse events cover pressing or releasing mouse-buttons and double-clicks. Especially

since the abstract event passing interface provides no guarantees on action execution, a mouse pointer placement and verification system had to be implemented. Such a system not only makes mouse movement independent of the original users movements, but also allows users to jump to any previous event with a defined mouse pointer state. State transitions are triggered either through the arrival of appropriate feedback from the runtime environment or through a time-out. Failures can happen either by mismatching the precondition or the postcondition. If the precondition is not met within a defined time-out, the system may try to step back until a previous precondition matches and retry event execution from that point. In the case of a mismatched postcondition, the system could check if the precondition still holds and retry the last event

Figure 6. Workflows and software components involved when accessing a digital artifact of a certain object type

execution. Although both recovery strategies may cover the most common failures, the operator still needs to decide which strategy is appropriate. The described approach is based on the GRATE system architecture using VNC for input/output abstraction.

6. STEP 6: PRESERVING NECESSARY SOFTWARE COMPONENTS

A major factor in the discussion of emulation strategies is widely missing. The needed additional software components are implicitly used but are not categorized and officially archived. Thus a missing operating system or firmware ROM of a home computer might render a digital object completely unusable, even with a perfectly running virtual replacement of the original machine. A first step to formalizing the access to digital objects of different types were view paths (see Step 3). They do not only define workflows to be implemented as mentioned in the last section but generate lists of needed additional software components.

Rendering digital artifacts requires, depending on the object type, a large and complex set of software components, not only the original software application and operating system. Other dependencies such as font sets, decompression software, codecs for audio and video files, and hardware drivers for video output, sound cards and peripheral devices must be met as well (Fig. 1). Typically, the more recent the environment, the higher the level of complexity and number of different components required. In addition to storing and handling the digital objects themselves, it is essential that we store and manage this complex set of software components (Fig. 6). These dependencies and requirements can be formalised using view paths (pathways) both for emulation and migration approaches to preservation [16]. Despite the considerable efforts on digital

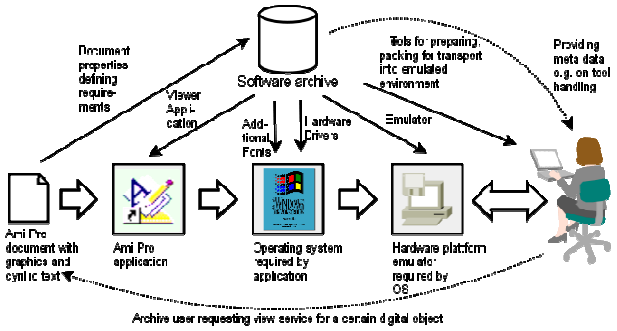


Figure 6. Workflows and software components involved when accessing a digital artifact of a certain object type

preservation research, this essential groundwork has until now been largely neglected. This could lead to fatal gaps in the preservation workflows of future generations.

Another scenario where a comprehensive and well-managed software archive is essential is when a memory institution receives the legacy of an important writer, scientist or politician. Typically such archives have not been actively managed, but are nonetheless of importance for cultural heritage. Depending on the age of the material, software archeology techniques may be required to provide access to this material. Established preservation organisations such as libraries and technical museums would be the natural providers of such a capability.

Legal Issues Alongside managing the software components and associated documentation, a software archive must tackle the legal and technical problems of software licensing. A reputable institution must abide by the licences associated with the software it uses. For proprietary software, this may severely limit the rights of the institution to use the software to provide preservation services. Furthermore, technical approaches to protecting intellectual property, such as Digital Rights Management (DRM), copy protection mechanisms, online update or registration requirements all create significant problems for a software archive. To tackle this problem will require the cooperation of software manufacturers with a designated software archiving institution, to provide suitably licensed unprotected copies of software for long-term preservation purposes. We recommend the development of an approach similar in concept to the legal deposit approach used by many national or copyright libraries.

7. STEP 7: PROVIDING REFERENCE ENVIRONMENTS

The emulator has to be run in the environment the archive user is working with. The user is to be enabled to construct a view path from the original object. The base platform for emulation should be chosen from the

most popular operating systems and computer hardware of a particular timespan. This prevents the establishment and costly operating of a hardware museum on one hand side and helps the user to orient him or herself more easily in a familiar surrounding. Additionally, the reference environment should offer easy access to all meta-data and required toolkits [16].

Every computer platform, historical as well as current, has its own complexities and concepts and most of the future computer users wont find old user interfaces as easy to use as we might think today. The same is true for set-up and installation routines of emulators and ancient operating systems. Another challenge arises from the transport of the requested artifact from the current into its original environment. Thus it would be desirable to automate the significant parts of the process in specialized departments of memory institutions with trained personnel and offer the services within a framework over the internet. This eases the complex procedures to be run on average computers and reduces the functionality to the viewer, e.g. in a web browser. The user gets the results presented via a virtual screen remotely on her or her computer (Fig. 4). With GRATE, a pilot was programmed to develop a prototype of an emulation service. This service is based on available open source emulators mentioned above and allows them to run on a remote basis.

Within the PLANETS framework [5], such emulation services can be integrated in more complex workflows of digital preservation. Emulated systems can be used as alternative endpoints of a migration workflow in order to allow an interactive view of the digital object in its original creation environment. Moreover, emulation itself could be used as a migration service in a different workflow. The PLANETS framework offers interfaces for web services for common tasks in digital preservation, like the characterization, validation, viewing, comparing, modifying and migrating of digital objects. Two PLANETS services are of particular interest for emulation.

The PLANETS *view* web services interface (Fig. 5) is designed to render a digital object. The service takes a digital object and returns a URI pointing to the rendered result. If the digital object requires a running rendering engine, the service offers methods for querying the engine's state and allows sending commands to it. The emulation viewing service offers access to already configured and ready-made emulators and software images. The web service accepts a list of digital objects and injects them into the running OS. The user is able to explore the environment, create, view or modify digital objects with their original application and compare the result visually with their appearance in current applications or migrated version of them.

By using the view interface for installing applications and their dependencies, not only can all steps of the recording procedure be recorded, but also might get annotated by the user. For each installation step this information is kept together with the system state before and after the installation, the application files and the system setup (e.g. which transportation option was used to provide the installation image) in the software archive. This way the software archive is able to:

- Calculate possible dependencies and view-paths for every known application setup;
- Ensure integrity of every view-path endpoint. This is achieved by keeping all intermediary setups and necessary installation files but most importantly the carried out installation steps;
- Calculate possible migration paths, provide access to the necessary files, the required set-up and the recording of the actual object migration.

The Planets Migrate Web service interface offers the ability to use various services to transform a digital object into a selected output format. The interface expects a digital object as input format and a designated output format accompanied with a list of service specific parameters. The outcome will be either a successfully transformed digital object or an error message. The *migration by emulation* services retrieves at instantiating time a so called view path-matrix from the software archive, which describes supported format migrations

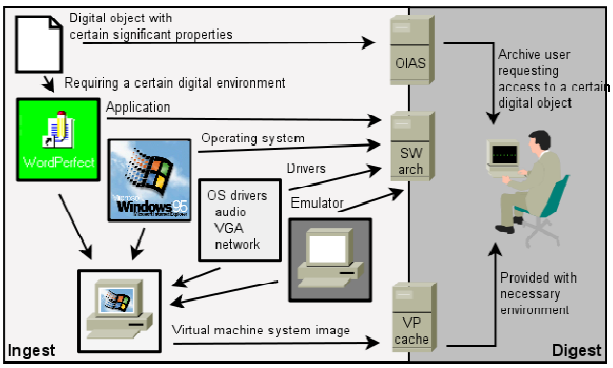


Figure 7. Future workflows to be implemented and integrated for emulation strategy

and then registers itself within the Planets framework. If the service is called with a supported view path, a view-path vector is requested from the software archive. This vector consists of a pointer to a system emulation engine, an appropriate runtime environment (e.g. a container file already set up with the appropriate operating system and applications) and a recorded interactive migration workflow. The digital object passed by the caller is injected into the runtime environment. After running the recorded workflow, the service returns all files (within a ZIP container) as digital object. Usually such a service is executed without visual control. However, for debugging and in case of an unrecoverable error, the view interface can be attached to the runtime environment.

8. CONCLUSION AND FUTURE WORK

Emulation is a very versatile and durable solution for retaining access to any kind of digital content. For some digital objects such as games, educational software or research applications, it is actually the only possible way as these objects usually can not be migrated. Nevertheless, emulation is not widely adapted to preservation frameworks and solutions in operation today for a number of reasons. There is still a trade-off between the well-established commercial virtualization tools without any long-term preservation focus like VMware and similar products, and preservation projects like Dioscuri and UVC are still missing major features to fullfill the average needs of memory institutions. UVC implements necessary web service interfaces for preservation frameworks but offers very limited support for different digital object types. Dioscori still lacks the support of newer Windows and other operating systems from Windows 95 on.

Emulators like QEMU, MESS or ScummVM prove the validity of the Open Source approach. Especially QEMU has reached a stage rendering it suitable to be integrated into preservation strategies utilizing emulation. However, there is a gap in the development focus between the developer community of emulators on the one side and the professional deployment in long-term archiving on the other. The developers of the above-mentioned emulators have different development goals than archiving organizations. The present state of quality assurance is far from satisfactory and must be extended by suitable, preferably automatic, test scenarios which verify all important aspects of correct CPU and hardware replication. Generally, the question remains if the development methods followed by QEMU or Dioscuri, which originated from particular development environments and paradigms, will be valid for a long-term timespan.

Long-term perspective If one wants to ensure sustainability, the future development should be actively supported by a suitable syndicate, such as a large archiving organization. It shows that such a project can only be carried out with wide support from personal and Open Source Communities and needs a long-term perspective. A long-term archiving strategy like emulation can not be achieved by a single organization, even of the size of a national library, since the specific knowledge of particular computing architectures and digital object types will be spread between the archive and science communities. Projects like the Open Planets Foundation,[7] which continues the Planets archiving framework, could serve as an example for new approaches.

Nevertheless, one still needs to understand how to operate an old computer environment. Today, many of us still remember older environments such as MS-DOS and early Windows, but soon even those experiences will be lost. Thus an emulation strategy has to be supplemented with means to preserve a more complete idea of past digital environments rather than just their hardware emulators and software components. Therefore, manuals, tutorials and other supporting documents need to be preserved and kept available as well. Tacit knowledge could be preserved e.g. in workflow recordings of those past environments. This produces the base to offer appropiate access environments to the future archive users. Automated workflows will play a major role for migration-by-emulation strategies and the set-up of past digital environments for the deployment on reference workstations (Fig. 7).

Especially in regard to the preservation of a wide know-how, a distributed approach should be chosen in which single institutions specialize on one area but an intensive exchange and shared access to the repositories remains possible. Especially when it concerns the preservation of various localized variants of software, cooperation of the national institutions is proposed. A particular requirement of software archiving lies in the preservation of specific components, like hardware drivers for the network, graphics or sound cards offered by the emulators. In addition to this are codecs or fonts that are required for particular types of videos, audio or documents.

[7]OPF, *http://www.openplanetsfoundation.org*

9. REFERENCES

[1] David Bearman. Reality and chimeras in the preservation of electronic records. *D-Lib Magazine*, 5(4), 1999.

[2] Christoph Becker, Hannes Kulovits, Michael Kraxner, Riccardo Gottardi, Andreas Rauber, and Randolph Welte. Adding quality-awareness to evaluate migration web-services and remote emulation for digital preservation. In *Proceedings of the 13th European Conference on Digital Libraries (ECDL09)*, 2009.

[3] Max Burnet and Bob Supnik. Preserving computing's past: Restoration and simulation. *Digital Technical Journal*, 8(3):23–38, 1996.

[4] Adam Farquhar and Helen Hockx-Yu. Planets: Integrated services for digital preservation. *International Journal of Digital Curation*, 2(2), 2007.

[5] Ross King, Rainer Schmidt, Andrew N. Jackson, Carl Wilson, and Fabian Steeg. The planets interoperability framework. In *Proceedings of the 13th European Conference on Digital Libraries (ECDL09)*, pages 425–428, 2009.

[6] Raymond Lorie. *The UVC: a Method for Preserving Digital Documents - Proof of Concept.* IBM Netherlands, Amsterdam, PO Box 90407, 2509 LK The Hague, The Netherlands, 2002.

[7] Commission on Preservation, Access, and The Research Libraries Group. Report of the taskforce on archiving of digital information. WWW document, *http://www.clir.org/pubs/reports/pub63watersgarrett.-pdf*, 1996.

[8] Klaus Rechert and Dirk von Suchodoletz. „Tackling the problem of complex interaction processes in emulation and migration strategies," *ERCIM News*, (80):22–23, 2010.

[9] Tristan Richardson. The rfb protocol. WWW document, *http://www.realvnc.com/docs/rfbproto.pdf, 2009*.

[10] Jeff Rothenberg. Ensuring the longevity of digital information. *Scientific American*, 272(1):42–47, 1995.

[11] Jeffrey van der Hoeven. Dioscuri: emulator for digital preservation. *D-Lib Magazine*, 13(11/12), 2007.

[12] J.R. van der Hoeven, R.J. van Diessen, and K. van der Meer. "Development of a universal virtual computer (uvc) for long-term preservation of digital objects," *Journal of Information Science*, 31(3):196–208, 2005.

[13] Raymond van Diessen and Johan F. Steenbakkers. *The Long-Term Preservation Study of the DNEP project - an overview of the results*. IBM Netherlands, Amsterdam, PO Box 90407, 2509 LK The Hague, The Netherlands, 2002.

[14] Remco Verdegem and Jeffrey van der Hoeven. Emulation: "To be or not to be," In *IS&T Conference on Archiving 2006, Ottawa, Canada, May 23-26*, pages 55–60, 2006.

[15] Dirk von Suchodoletz. *Funktionale Langzeitarchivierung digitaler Objekte – Erfolgsbedingungen für den Einsatz von Emulationsstrategien*. Cuvillier Verlag Göttingen, 2009.

[16] Dirk von Suchodoletz and Jeffrey van der Hoeven. Emulation: From digital artefact to remotely rendered environments. *International Journal of Digital Curation*, 4, 2009.

[17] Nickolai Zeldovich and Ramesh Chandra. Interactive performance measurement with vncplay. In ATEC '05: *Proceedings of the annual conference on USENIX Annual Technical Conference*, pages 54–64, Berkeley, CA, USA, 2005. USENIX Association.

BABS2: A NEW PHASE, A NEW PERSPECTIVE IN DIGITAL LONG-TERM PRESERVATION – AN EXPERIENCE REPORT FROM THE BAVARIAN STATE LIBRARY

Tobias Beinert **Markus Brantl** **Anna Kugler**

Bavarian State Library
Ludwigstraße 16
D-80539 München
surname@bsb-muenchen.de

ABSTRACT

BABS is an acronym for Library Archiving and Access System (Bibliothekarisches Archivierungs- und Bereitstellungssystem), which constitutes the infrastructure for digital long-term preservation at the Bavarian State Library (BSB). During the two-year project BABS2 funded by German Research Association (DFG) BSB focuses together with the Leibniz-Supercomputing Centre (LRZ) on advancing its organizational and technical processes under the aspect of trustworthiness according to the nestor criteria catalogue. Important achievements are e.g. framing an institutional policy for digital preservation including local, regional, national tasks of a large-scale research and archive library, conducting and evaluating a survey concerning the archiving requirements of all BSB departments, documenting the ongoing archiving processes, introducing an appropriate quality management and improving the scalability of the preservation system. Additionally BSB participates in different national and international committees.

This experience report sheds light on the various organizational and technical aspects which have to be taken into consideration when enhancing an existing infrastructure for digital long-term preservation.

1. INTRODUCTION

Today the Bavarian State Library (BSB)[1] as universal and international research library manages the largest digital archive for cultural heritage in Germany, now containing more than 380 million files with a total amount of 218 terabyte (June 2010). Several mass digitization projects, including the public private partnership with Google, as well as different web archiving and audio digitization projects contribute to the considerable scale and variety of resources in the archive. Not only these huge dimensions, but also new political responsibilities (e.g. the inclusion of online publications of public authorities in the legal deposit law) require several organizational and technical enhancements of the existing archiving infrastructure.

Responsible for long-term preservation within BSB is the Munich Digitization Center/ Digital Library (MDZ)[2], which during the last years has set up a "Library Archiving and Access System" (Bibliothekarisches Archivierungs- und Bereitstellungssystem, BABS[3]) in collaboration with its strategic partner, the Leibniz Supercomputing Centre[4]. Current preservation responsibilities of BSB include:

- Digitized books produced by BSB or by commercial partners
- Born-digital documents delivered according to the current legal deposit act (e. g. governmental publications of the Bavarian state and of other German governmental institutions)
- Digital resources (e. g. websites, open access publications) belonging to the virtual library of BSB's special collection fields

as well as all other electronic media produced or licensed for use by BSB.

For digitized materials the MDZ has developed, implemented and optimized a production line for all the relevant processes, e.g. preparation, scanning, metadata enrichment, delivery, and archival storage among others. A self-developed software-tool (so-called ZEND) fosters this workflow and makes it possible to cope with mass digitization and preservation of a wide range of materials from medieval manuscripts and incunabula, to journals and newspapers, as well as photographs and audio documents [1].

[1] www.bsb-muenchen.de

[2] www.digital-collections.de
[3] www.babs-muenchen.de
[4] www.lrz-muenchen.de

2. THE PROJECT BABS2

The current project BABS2, funded by the German Research Association (DFG)[5], faces the challenge to consolidate and improve the existing architecture for digital long-term preservation, integrate it into the overall organization of the library and adjust it to newly upcoming requirements. The aim is to build a trustworthy and scalable digital archive as part of a national network for digital preservation.

The ongoing improvement processes cover organizational aspects such as designing a digital preservation policy for the BSB, developing new workflows, documenting the existing workflows, as well as technical aspects such as re-structuring the present storage system and introducing (periodic) virus- and checksum scans.

During the project, experiences with innovative methods (e.g. preservation planning, self-auditing based on criteria for trustworthiness) will be made in the fields of organization, evaluation and improvement of digital preservation. In accordance with nestor[6] and in collaboration with the German National Library (Deutsche Nationalbibliothek, DNB)[7] and regional libraries, models for national cooperation will be developed.

3. ORGANIZATIONAL ENHANCEMENTS

3.1. Consolidation of the digital archive

The digital archive of the BSB was established out of the need to store the rapidly growing amount of data beginning with the first digitization projects in 1997. Now, more than ten years later, further consolidation of the organizational and technical infrastructure inside the overall institutional framework of the library is necessary.

A long-term-preservation unit inside the Munich Digitization Center / Digital Library was established already in 1999. With regard to the changing organizational structure of the BSB its tasks and responsibilities were further clarified. At present the long-term preservation unit is responsible for the connection to the digital production as well as for research and development.

A first milestone of the BABS2 project was the design of a policy which clearly defines the aims of the digital archive of the Bavarian State Library as one of the most important cultural heritage institutions in Germany, as the archive library of Bavaria, and as the head of the Bavarian Library Network. Besides a concise mission statement, it states the reasons for BSB's responsibility in the field of long-term preservation and tries to shape a basic profile for its collection and archiving duties in the digital world. Furthermore it comprises an explanation of the general principles which BSB adheres to (e.g. provision of customer-oriented digital services; ensuring trustworthiness; long-term preservation as a cooperative business etc.). The existence of such a policy is itself also one main criterion of trustworthiness according to the nestor criteria catalogue [4]. A first draft has been completed, it is now up for discussion by the responsible departments of the library and has yet to be adopted officially by BSB's head office.

The preparation of a written mutual agreement between BSB and LRZ in digital long term preservation formed a next important task for organizational consolidation in the BABS project. Since 2004 both institutions have been working together in several projects. Building up a jointly operated technical infrastructure and transferring into routine business was a central milestone in digital long term preservation for both sides. In the context of the BABS2 project a refinement of the Service Level Agreements (SLAs) between BSB and LRZ takes place.

A further consolidation step of BSB's digital archive was the examination of the existing archiving workflows, including those for the ingest of monographs and periodicals as well as the workflow for digital materials and their abstraction as process models. In addition to a detailed documentation of all current processes and activities, which provides the basis for trustworthiness, we developed new workflows e.g. for legal deposit and web archiving.

To prepare for future tasks in the library and the Bavarian Library Network we are conducting an inventory survey/stakeholder analysis for all major departments of the library and the partners of the network. Our aim is to review the digital material available and to detect present and future requirements for digital long-term preservation. The results will lead to the design of adequate organizational and business models for long-term preservation at the BSB and the Bavarian Library Network.

As part of our quality management we organized a workshop together with our project partners from LRZ, in order to self-evaluate our preservation architecture. In a first step we reviewed our digital archive according to the concept of trustworthiness set out in the nestor-criteria catalogue [2] using the following assessment scale:

- 1 = conception
- 2 = in process of implementation
- 3 = completely fulfilled

Many of the criteria were well fulfilled, such as all actions are based on legal and contractual regulations (e.g. controlled access to the digital documents) or the definition of necessary metadata, but other criteria concerning the organizational structure of the digital

[5] German Research Association (Deutsche Forschungsgemeinschaft): www.dfg.de
[6] www.langzeitarchivierung.de
[7] www.d-nb.de

archive (e.g. distributed responsibilities over different departments and institutions) need further enhancement.

In a second step we applied the self-audit method of DRAMBORA (Digital Repository Audit Method Based on Risk Management)[8] to assess our digital archive. As we especially wanted to review the transfer of our digitized images to the storage at LRZ we focussed on a risk assessment of integrity and authenticity in the areas of ingest and storage. The risks we could identify in this area are in great parts already covered quite well by LRZ's own institutional risk management programme, but now have to be specified in greater depth for our joint preservation activities.

3.2. Cooperation activities

The long-term-preservation unit of BSB takes part in national and international collaborations, e.g. in committees such as the German competence network for digital preservation nestor and standardization working groups at DIN/ISO.

Within the framework of nestor[9] and in collaboration with the DNB and regional libraries, BSB contributes its share to the very challenging task of developing cooperative models for long-term preservation in the federal state of Germany. The activities in this area include amongst other things the planning of exchanging information packages in both directions between the BABS-system and DNB's system kopal as well as the participation in the LuKII -project[10] which aims at setting up a LOCKSS-network for Germany and testing the interoperability of that network with repositories and archival systems. With the Library of Congress BSB established a private LOCKSS network in order to test the exchange of electronic official publications.

Furthermore BSB is actively involved in the working group for a National Hosting Strategy within the framework of the Priority Initiative "Digital Information"[11] by the Alliance of German Science Organisations. A first result of these activities was the publication of a final report *Ensuring Perpetual Access* by Charles Beagrie Limited on the establishment of a federated strategy on enduring access and hosting of digital resources for Germany in March 2010 [2].

4. TECHNICAL ENHANCEMENTS

4.1. Enhanced AIPs, integrity and authenticity

For digitized books up to now bibliographical and structural metadata is saved with the digital objects. Due to new requirements regarding the heterogeneity of digital resources of a universal library (manuscripts, rare books, special collections) and the growing complexity in the technical production we decided to store further technical information for the singular image. We have developed an additional workflow to generate and save further preservation metadata (technical as well as event metadata) on the image level according to the PREMIS standard.

The technical metadata is extracted with jhove, but we are also evaluating the possibility of using FITS[12], because this tool integrates different extraction and validation services which comply better to our requirements. FITS is in comparison to jhove e.g. able to extract information about the ICC colour profile, which we need for further possible migration actions. Event metadata is saved in the process of ingest (e.g. creation, validation results, enrichment, normalization/migration actions). The technical metadata as well as the event metadata is stored in xml-files next to the digital object.

In a parallel effort the structural metadata of our digital objects is revised, so we will be able to tie an enhanced and enriched AIP which includes explicit metadata for long-term preservation.

Our aim is to perform virus checks and scans of the generated checksums at certain points of the established archiving workflow and store this information inside the AIP. We tested the required time and processor performance of these checks, which showed that generating as well as scanning a certain amount of checksums can be included into our digitization workflow without interfering with the on-going processes. We identified the critical stages of the archiving workflow and fixed certain points of time where the checksums can be generated and accordingly checksum and virus scans can be performed. All information is stored in the newly designed AIP of our digitized books. In terms of risk management we are developing concepts on how to proceed in case of an integrity violation (see 3.1.).

4.2. Preservation planning

In a joint workshop with the TU Vienna in 2009 we designed a preservation plan for format migration of a selected collection of digitized books (16th Century Printings) using the PLATO tool which supports the process of decision-finding and documentation [3]. We considered the option of migrating the selected digitized images from TIFF to JPEG 2000 as new archival format. Following the planning workflow of PLATO we first of all defined our requirements according to the digitization standards and preservation policy of the BSB. Some of our main requirements were e.g. to keep the resolution and the ICC colour profile of the image after the migration, to reduce storage costs or to allow the creation of full-text (OCR).

[8] www.repositoryaudit.eu

[9] www.langzeitarchivierung.de

[10] http://www.ibi.hu-berlin.de/forschung/digibib/forschung/projekte/LuKII

[11] http://www.allianzinitiative.de/en/core_activities/national_hosting_strategy/working_group/

[12] http://code.google.com/p/fits/

In a 2nd step we tested different alternatives of migrating from TIFF to JPEG 2000 with several open source tools, which showed different outputs. Finally we evaluated the results and built the preservation plan for our collection: the alternative of "keep status quo" excelled over the other possibilities and was thus our recommended preservation action. According to our changing requirements and the development of new or improved tools we need to review our preservation plan on a regular basis.

During this workshop we gained the necessary methodical skills to pursue further preservation planning for other collections, as for example for the legal deposit which will focus on pdf to pdf/a migration.

4.3. Scalability

In the BABS2 project we cooperate with the Leibniz Supercomputing Centre to test the scalability of our digital archive. Tests with different storage management systems (SAM/QFS, TSM/HSM) showed that the huge amount of digital data BSB produces can't be handled easily by the well-established software. Together with previous scalability experiences (e.g. migration of storage media, handling the data of the Google-project [5]) it became obvious that a complete re-structuring of the storage system by virtual units should be done in order to allow further growth, improve access and performance, as well as allow migration of storage media which does not interfere with the daily routine.

4.4. Perspective: Introducing a new technical solution for preservation

Due to risen requirements, large scale of data, diversity of resources and a broader archiving focus, it is necessary to introduce a new, more robust system with long-term preservation functionalities. The above described experience and the gained knowledge prepared us for the introduction of the new archiving system "Rosetta".

The Digital Preservation System "Rosetta" was developed by ExLibris in a partnership with the National Library of New Zealand. Rosetta will enhance the technical infrastructure and its associated preservation workflows at BSB. The existing workflows will be improved, unified and consolidated into one single system. By introducing Rosetta, BSB is on the way to apply several new features regarding long-term preservation. It offers e.g. a detailed risk analysis for each file, which provides the basis for the new preservation planning module. Preservation actions can be performed on a selected set of files according to the beforehand defined preservation plan.

The open platform architecture of the new system allows an easy interconnection with different external systems via customized API/SDK developments. All these efforts and activities improve the OAIS compliancy, the trustworthiness, scalability and robustness of BSB's long-term preservation activities substantially.

The introduction of Rosetta starts with a pilot phase. During this time the specifications for the transition of three designated workflows into the system are designed, implemented and tested:

- Digitized objects
- Legal deposit
- Webarchives

According to these workflows several external systems and tools need to be integrated and tested. At the end of this phase Rosetta should be switched over to routine business and gradually the other workflows should be adapted to Rosetta. The already archived objects in the existing architecture will be migrated to the new archival system step by step.

In a second stage after introduction, libraries out of the Bavarian Library Network will join the BSB in using the system.

5. REFERENCES

[1] Brantl M., Schoger A.: „Das Münchener Digitalisierungszentrum zwischen Produktion und Innovation", *Information - Innovation – Inspiration The Bavarian State Library. 450th Anniversary,* 2008.

[2] Charles Beagrie Limited in association with Globale Informationstechnik GmbH: Ensuring Perpetual Access: Establishing a Federated Strategy on Perpetual Access and Hosting of Electronic Resources for Germany. http://www.allianzinitiative .de/fileadmin/hosting_studie_e.pdf

[3] Kugler, A., Brantl M., Beinert, T., Schoger A., Kulovits H., Rauber, A. (2009): "From TIFF to JPEG 2000? – Preservation Planning at the Bavarian State Library Using a Collection of Digitized 16th Century Printings", *D-Lib Magazine*, Vol.15, 2009. http://www.dlib.org/dlib/november 09/authors/11authors.html

[4] nestor working group for trusted repositories and certification (2009): nestor criteria – catalogue of criteria for trusted digital repositories, Version 2. http://nbn-resolving.de/urn:nbn:de:0008-2010030806

[5] Wolf-Klostermann, T. "How to cope with 300.000 scans a day. Managing large scale digital collections in practice – the Bavarian State Library and the Leibniz Supercomputing Centre approach the next level of mass digitisation", *Archiving 2008,* 2008.

PERSONAL ARCHIVING: STRIKING A BALANCE TO REACH THE PUBLIC

William G. LeFurgy
Library of Congress
101 Independence Ave, SE
Washington, DC 20007

ABSTRACT

Never have so many people documented so much about their lives. Digital technology has empowered individuals to build large, rich collections of photographs, videos, e-mail, documents and other information. But the ability to create digital content is far outstripping personal capacity to manage and keep it over time. Looking ahead over the next decade, it is possible to foresee two consequences for libraries and archives. The most obvious and certain is that digital accessions of personal materials will supplement, and eventually surpass, traditional analog materials. Another outcome is more subtle and speculative: people seeking trusted guidance about how best to manage their important digital items. Memory organizations are reasonable places for people to go in search of such guidance, and this presents an opportunity to provide a valuable—and highly visible—public service. The Library of Congress National Digital Information Infrastructure and Preservation Program is undertaking a project to provide guidance aimed at the general public in connection with personal digital archiving. The project focuses on interacting with people through several different channels, including web-based written instructions, video productions, and social media. The Library is also exploring use of public events such as "Personal Archiving Day" to engage directly with people. In developing a strategy for this program, the Library has to balance professional practice with the need to clearly communicate with non-specialists in a Web 2.0 environment.

1. RISING TIDE OF PERSONAL DIGITAL INFORMATION

Residents of the developed world are generating an astonishing amount of personal digital information. Reliable figures are hard to come by, but the Twitter archive is estimated to consist of five terabytes1; Flickr has and estimated 4.3 billion pictures and Facebook may have anywhere from 15-60 billion pictures[2]. Millions of digital cameras (and phones with cameras) are in circulation. Just one of many computer manufacturers expects to sell 25 million personal computers itself over the next year[3]. An information technology market analysis claims that 70 percent—or about 880 petabytes--of the annual "digital universe" is generated by personal users[4]. However the phenomenon is looked at, it is clear that 1) a typical consumer has accumulated a staggering quantity of data, and 2) the trend line is headed up, probably dramatically so.

At the same time, it appears that categories of newly generated analog personal information are in steep decline. Kodak, an iconographic analog stalwart, has struggled in recent years, and one business analyst recently forecast more trouble for the firm due to its "exposure to the secular decline in analog film." The company has, since 2004, pinned its hopes on "digital photography services and printers and away from photographic film[5]." The U.S. Postal Service also faces economic challenges that stem in part from "digital alternatives such as electronic bill payment [and] e-mail document delivery[6]." Home movies on 8mm film and other analog formats have given way to digital video recorders.

It seems that many people are putting all their personal information eggs in a virtual digital basket. This is, of course, a risky strategy. A shoebox filled with photographic prints, letters, home movies and the like can easily last for years with minimal care and easily pass from one generation to the next. Personal digital content presents a whole new challenge. It is often scattered across a variety of websites, devices and storage media. Content is frequently disorganized and

[1] Library Journal
http://www.libraryjournal.com/article/CA6726233.html

[2] http://www.personalarchiving.com/2010/02/conference-notes/

[3] http://www.google.com/hostednews/ap/article/ALeqM5iEIzDh7j2jM9lXEuW7TFxLWKmrlQD9FKL1H80

[4] http://www.emc.com/collateral/demos/microsites/idc-digital-universe/iview.htm

[5] http://www.reuters.com/article/idUSN2923844820100429

[6] http://articles.sfgate.com/2010-04-19/news/20855490_1_postmaster-general-john-potter-postal-service-mail

subject to spotty strategies for selection, replication and metadata. Commercial services are frequently used to store personal content despite the fact that such services are under no long-term obligation to keep or provide access to data. And even if a user keeps a copy of their data, digital media at this point are fundamentally non-archival.

2. CONSEQUENCES FOR MEMORY ORGANIZATIONS: EXTENDING CURATORIAL PRACTICE

Looking ahead over the next decade, it is possible to foresee two consequences in connection with personal digital information for libraries, archives and other memory organizations. The most obvious and certain is that digital accessions of personal materials will supplement, and eventually surpass, personal analog materials. Salman Rushdie's donation of his personal papers recently received attention because they included "four Apple computers (one ruined by a spilled Coke) [and]18 gigabytes of data[7]." But there will come a day when such donations are routine, as the generation that first adapted to digital technology start offering the fruits of their labors to collecting institutions. This assumes, of course, that the fruits are preserved in the first place. There is plenty of room for concern that e-mail correspondence, for example, will disappear into the ether[8]. But it would seem that in the future, where there are materials to donate, they will be increasingly digital.

Given this, it makes sense for memory organizations to develop a strategy for dealing with the situation. Actually, there might parallel strategies. One could focus on applying traditional curatorial approaches to prospective digital collections. While more work is needed, some attention has been devoted to this area in connection with ideas about Trusted Digital Repositories and other approaches. It does remain to be seen just how ready most institutions—even the biggest—are for adapting their practices to bringing in personal digital collections.

Another strategy needs to focus on first understanding how personal collections of scholarly interest are built and managed, and then exploring how to provide guidance, tools and services to help creators build viable personal digital archives. This constitutes a more proactive role prior to making institutional stewardship arrangements, and would seek to extend curatorial action or influence in a pre-custodial (or is it post-custodial?) manner. The Digital Lives Research Project at the British Library is a premier example of such an effort[9].

[7] http://www.nytimes.com/2010/03/16/books/16archive.html

[8] http://www.nytimes.com/2005/09/04/books/review/04DONADIO.html?pagewanted=all

[9] http://www.bl.uk/digital-lives/

3. CONSEQUENCES FOR MEMORY ORGANIZATIONS: PROVIDING BROAD-BASED GUIDANCE

The other big consequence looming as a result of the growth of personal digital holdings is that millions of people are going to need advice about how to save important parts of their collections and pass them on to family members or other interested parties. Memory organizations are reasonable places for people to go in search of such guidance, and this presents an opportunity to provide a valuable—and highly visible—public service.

This is a tricky business for memory organizations, however. Curators are by definition experts with deep and arcane knowledge. They focus on material that is of scholarly interest, which typically is a mere fraction of the larger information universe. To the extent that advances have been made in digital preservation and curation, they are tightly bound with specialized curatorial and technical concerns relating to complex issues relating to authenticity, metadata, validation and verification and fixity. While strides have been made in generalizing and simplifying some digital preservation methods, much current practice remains institution specific and opaque to the average person.

There is a further complication to providing archiving guidance to non-specialists. Web 2.0 has brought about a hunger for information that is quickly found, read and understood. Many people have limited patience for carefully nuanced, specialized information sources. When writing for the web, authors are given advice to write text that is easily scanned, as opposed to read word by word. This means bullets and many fewer words than conventional writing.

If memory organizations want to meet the challenge of providing broad-based digital archiving guidance they need to identify the bare bones of good practice, seek out effective channels to present information and work to engage users in different ways. And while much of this work will take place through computer mediated methods, they is very much a place for meeting with people directly to exchange information.

4. THE LIBRARY OF CONGRESS AND PERSONAL DIGITAL ARCHIVING

The Library of Congress National Digital Information Infrastructure and Preservation Program is undertaking a project to provide guidance aimed at the general public in connection with personal digital archiving. The project focuses on interacting with people through several different channels, including web-based written instructions, video productions, and social media. The Library is also exploring use of public events such as "Personal Archiving Day" to engage directly with people. In developing a strategy for this program, the Library has to balance professional practice with the

need to clearly communicate with non-specialists in a Web 2.0 environment.

In May 2010, NDIIPP posted new and expanded guidance for personal digital archiving on the program website[10]. The information focus on six categories: digital photographs, digital video, digital audio, e-mail, personal digital documents and websites and social media. Each category is built on a basic structure of advice that is distilled from professional practice. Specific issues relating to a particular category of personal digital information are highlighted. The advice is described as "basic and is meant to be a place to get started"—it makes no claim to be one-stop shopping for everything that an individual needs to know for digital preservation.

At this stage, the NDIIPP personal digital archiving guidance is regarded as a "beta": it should work as intended but future enhancements are expected, if not required. The program will conduct user testing and will seek comments about how to improve the guidance while adhering to its intentional Web 2.0 presentation. The current guidance framework is built on the following components:

- Identify the full scope of your collection
- Decide which parts of it you want to save
- Organize and describe what you selected
- Make copies and store them in different places

A fifth component, export selected items from individual programs and services, is used when discussing e-mail and other content with specific dependencies.

NDIIPP has also developed a Digital Preservation Video Series to convey information in a YouTube-friendly format[11]. The videos are, in fact, posted on YouTube as well[12]. They deal with a number of digital preservation issues and are meant to be engaging and informative. Individual videos are planned for each of the six content categories covered in the personal archiving guidance.

Also in May 2010, NDIIPP launched a Facebook page to engage with the public about digital preservation[13]. The intent is to use the page as a separate channel for distributing information and most especially as a way to interact and with interested people and field specific questions and concerns. Despite the newness of the page, the Library is impressed with how rapidly people are "liking" it; expectations are that the page will play a key role in bringing NDIIPP to many more individuals than ever before.

Despite the pull of Web 2.0, the Library learned that place is still important—particularly a place for people to come and deal directly with curators and other experts about preservation practices. On May 10, 2010, the Library held its first "Personal Archiving Day" for the public[14]. About 200 people came to the James Madison Building on Capitol Hill in Washington, DC, to attend the event. Library staff gave brief talks on steps people can take to save their digital—and non-digital—information and staff also were available at content-specific tables to answer questions and talk about preservation issues. The event was held in conjunction with the American Library Association's inaugural "Preservation Week[15]." Interest shown in the work of ALA and the Library in connection with digital preservation is a solid indication that the public is eager for trustworthy advice—especially advice that flows through the right channels.

[10] http://www.digitalpreservation.gov/you/

[11] http://www.youtube.com/user/LibraryOfCongress

[12] http://www.facebook.com/#!/digitalpreservation

[13] http://www.digitalpreservation.gov/news/events/presweek2010/index.html

[14] http://www.digitalpreservation.gov/news/events/presweek2010/index.html

[15] http://www.ala.org/ala/mgrps/divs/alcts/confevents/preswk/index.cfm

SHERWOOD ARCHIVE PROJECT: PRESERVING THE PRIVATE RECORDS OF PUBLIC INTEREST

David A. Kirsch **Sam Meister**

Digital Archive of the Birth of the Dot Com Era Project
4556 Van Munching Hall
University of Maryland
College Park, MD 20742

ABSTRACT

Digital business records are more at risk now than ever before. The dynamics of entrepreneurial business, the fear of litigation and e-discovery, and the narrowing demands of shareholder capitalism compound the technological threats to the record of business. The Sherwood Archive Project, a project of the Digital Archive of the Birth of the Dot Com Era, seeks to mitigate the risks to business records by partnering with Sherwood Partners, Inc to develop strategies and workflows to preserve the records of failed business firms

1. INTRODUCTION

The following paper outlines current efforts of the Digital Archive of the Birth of the Dot Com Era project to preserve the digital records of business. A brief description of the nature of risk in relation to the record of business is given to portray the immediacy of the need for preservation efforts in this realm. The Sherwood Archive Project is outlined, including a description of work completed, future project stages, and illustration of the complexity of challenges inherent in attempting to preserve private records that are of public interest.

2. THE RECORD OF BUSINESS AT RISK

Why are business records more at risk now than in the past? This section summarizes ideas set forth more fully in Kirsch (2009) [4]. Information technology is implicated at every step, but technological change is not the only cause of the threat. The sources of the problem include entrepreneurship, litigiousness, and shareholder capitalism itself, each of which are indirectly affected by changes in the underlying technological landscape.

In the early growth stages of a typical entrepreneurial venture the focus is on searching for financing, customers, suppliers, employees, and above all, profit. Developing a records management program tends to be far down the list of priorities, and only if a venture survives and reaches maturity is a firm likely to create and implement policies in this area. This scenario is likely to persist into the near future and as entrepreneurship increases the allocation of resources for preservation of records will remain minimal.

The nature of shareholder capitalism also does not bode well for the survival of business records. In the traditional model of business archives, such as the Hagley Museum in Delaware or the Baker Library at Harvard, paper records were inadvertently or accidentally saved by the producing organization and then donated to an external archival institution when the private value of the records became lower than the public interest in those documents. The mandate of the shareholder context confronts the traditional business archive model by asserting that if a proposed action does not directly benefit shareholders, then that action should not be supported by responsible management. In such an environment, where the private benefits of contributing to an archive are questionable, it is expected that such donations will continue to decrease.

Lastly, the culture of litigiousness has direct repercussions that increasingly put the record of business at risk. The costs of legal discovery are unknown and therefore especially frightening to corporate entities attempting to predict and control every aspect of economic activity taking place within the boundaries of the firm. While statistics show that liability risk of a given record is low, the perceived threat of the cost of discovery has led to the development of strict record retention policies. Created to be in compliance with contemporary legislation the purpose of these policies is clear: every record produced within the boundaries of the organization should be saved until a certain date and then destroyed. These policies do not envision any records surviving beyond the organization.

Digital technology has produced the tools and mechanisms that allow information to be managed,

tracked, and destroyed with increasing efficiency. Companies can monitor the value of their digital assets, and in turn, swiftly delete these same assets when their value has decreased beyond the point of continued retention. No longer are records haphazardly kept and rediscovered later in an archival collection. The digital revolution, consistent with changes in the structure of the economy and the clearest dictates of shareholder capitalism, will destroy the Record of Business not by accident, but on purpose by making it manageable and valuable for a discrete, but limited period of time beyond which its value to the corporation falls below the cost to maintain it.

3. SHERWOOD ARCHIVE PROJECT

Against this background, the Sherwood Archive Project (SAP) represents an attempt to save the records of business by investigating the potential to preserve the "abandoned" records of failed companies. The SAP is one of the current efforts of the Digital Archive of the Birth of the Dot Com Era (DCA). Since 2002, the DCA has sought to identify, collect, and preserve a representative collection of born-digital records and related digital ephemera from companies that sought to exploit the commercialization of the Internet during the 1990s. Through previous projects and the resulting collections of the Business Plan Archive, the Dot Com Archive, and the Brobeck Closed Archive, the DCA has explored the complex privacy and confidentiality concerns that are associated with attempting to preserve digital business records. The SAP continues the efforts of these previous projects in confronting these challenges.

3.1. Background

In 2008, the DCA began a partnership with Sherwood Partners, Inc., a consulting firm located in Mountain View, California. Sherwood Partners provides a highly specialized service to the venture ecosystem. As the population of venture capital-backed startups expanded in the course of the 1990s, so too did the number of failed ventures, requiring venture investors to spend valuable time winding down old companies when they (and their LPs) would have rather focused on investing in new ones. Sherwood helped solve this problem by developing a novel "workout" mechanism. They have taken advantage of a legal code "Assignments for the Benefit of Creditors" – a state-based alternative to Chapter 7 Bankruptcy filing – that is available in many of the major states that venture capital backed companies either work in or have incorporated in. Sherwood has developed the Operating Assignment for Benefit of Creditors or ABC to better work in the venture capital community. Venture capitalists holding controlling stakes in failing startups "assign" all of the assets of the failing company to Sherwood, and in exchange Sherwood receives a fixed fee and/or a share of the total assets recovered in liquidation. In this way, venture investors outsource responsibility for the workout and winding down process to specialized professionals, simultaneously maximizing financial recovery, freeing the investors to look forward, and limiting the risk of potential entanglements resulting from public bankruptcy filings. Over the course of the past decade, Sherwood has served as the Assignee for several hundred failed firms, in the process returning tens of millions of dollars to creditors and investors. The records of these assignments contain distillations of billions of dollars spent in pursuit of uncertain opportunities, and their preservation promises to yield answers to many questions of immediate and historic interest. Functionally "abandoned" by their previous creators and owners, the records collected by Sherwood represent a valuable opportunity to both preserve these specific at risk business records of historic interest, as well as to determine the feasibility of developing preservation solutions for at risk business records in other similar contexts.

The Operating ABC is an elegant legal solution to the problem of how to efficiently liberate scarce financial, IP and human resources entangled in failing technology ventures so that they can be redeployed elsewhere in the entrepreneurial economy. However, the day-to-day workings of this process are more complicated. Different stakeholders hold differing views about the desirability and timing of initiating an Operating ABC. As a result, the transfer of control from the failing firm to the workout partner, in this case Sherwood Partners, is uncertain and tumultuous and presents unique preservation challenges.

3.2. Objectives and Methodology

The main objective of the SAP is to develop and implement a records management workflow for the paper and digital records collected as part of Sherwood's business processes. The end goal of this workflow is the transfer of selected records to an external repository for long-term preservation. In relation to the digital preservation lifecycle in OAIS [3] terms, the SAP is focused on investigating and producing solutions to challenges encountered during the pre-ingest time period. The conceptual nature of OAIS model does not specify the potential complexity that may be encountered during the pre-ingest stage, while in reality we see that many challenges arise. This context may be unique in the extent of the problem of business records, but may be that this problem is underappreciated in other contexts as well.

Early stages of the SAP focused on seeking understanding of the context for capturing records within the Sherwood workout process. A qualitative methodology for collecting data was employed during

these stages that included surveys, interviews, and field observations. A survey of the paper record collections previously collected by Sherwood was conducted to better understand the existing selection criteria for records being implemented during the ABC process. Interviews with key Sherwood staff members were conducted to gather data on the existing workflows carried out during ABC process, with specific attention paid to records selection, collection, storage, and disposal. Field observations, in the form of site visits to the facilities of failed firms undergoing workouts are ongoing and have provided valuable insight into the timeframe and context for records capture within a typical workout. During a field observation project staff accompany Sherwood staff to the former company facility soon after the ABC process has begun. Project staff will observe and record details on the site itself as well as the work being carried by Sherwood staff and associates.

Two main outcomes resulted from the data collected during the initial stages. First, a Selection Criteria for Paper Records was produced based on understanding of the existing Sherwood record selection criteria as well as the results of a review of the literature from the archives and records management fields focusing on the selection and preservation of business records. Second, a set of Policies and Procedures for Paper Records was developed to integrate the existing Sherwood staff workflows with the new Selection Criteria and additional steps for managing records. These two documents will assist Sherwood staff in selecting and managing records during their retention and use by Sherwood in the ABC process.

The current focus of the project is utilizing the same qualitative mechanisms to collect data on the existing digital records collections and the workflows carried out by Sherwood staff in capturing digital records during the ABC process. Outcomes of these current stages will include:

- Selection Criteria for Digital Records
- Recommendations for a workflow for the selection, capture, and transfer of digital records to external repository

3.3. Initial Results

Initial results of the early and ongoing stages of the project include increased understanding of the environment for potentially capturing records during the Sherwood workout process. This environment can be characterized as highly variable in relation to at least two important factors: time and access. First, the length of time that Sherwood staff will have to close down a client's former facility can vary from a few days to multiple months. This ever-changing window of opportunity to locate, identify, select, and capture digital records creates a challenging setting for digital preservation efforts. Second, the ability to begin the process of location, identification, selection, and capture may be impeded by multiple access obstacles. Digital records may reside on hardware that is no longer operational, on proprietary networks, or behind levels of password-protected encryption. Former staff with relevant information technology knowledge may be unwilling to assist in accessing digital records.

These factors are not necessarily new to those familiar with digital preservation efforts. Data located on encrypted devices or outdated media is a familiar challenge. Time is a well-known enemy to digital data. Through the process investigating the context for capturing records at the end of a company's lifespan we have discovered an additional important factor: people. Companies are organizations made of people, people who have the knowledge to assist in locating and capturing digital records. As the SAP moves forward, any new workflows for capturing the records of failed companies will likely need to incorporate agreements for cooperation with key personnel from those companies.

3.4. Challenges

The Sherwood Archive Project presents a distinct set of challenges related to the management and preservation of digital records.

In particular the issue of rights management in the digital preservation process presents a key area of concern in the SAP. In addition to the technical obstacles that must be confronted in any digital preservation scenario, such as file formats, complex digital objects, and encryption, the issue of rights management compounds the complexity of preserving digital records in the SAP. The ambiguous nature of this rights setting for digital business records engages multiple stakeholders with competing interests. Venture capital investors, who originally provide the funding to start firms, have a desire to maintain control over those firms and limit the potential of lawsuits resulting from discovery. Managers and founders of firms may be concerned about being tied to responsibility for the failure of companies and may seek to retain control of records. Employees are likely to have privacy and confidentiality concerns related to personal information embedded within corporate records.

Many of these potential concerns will be dealt with through the development of legal agreements with Sherwood and any external repositories. In these agreements we will incorporate the recommendations of previous projects [1] that have investigated the rights management challenges inherent in digital preservation. While determining the details of the legal agreements may be a lengthy and challenging process, the nature of the ABC process in assigning the assets of failed companies to Sherwood provides a fairly clear understanding of Sherwood as legal rights owners. However, a clear legal framework does not necessarily

resolve underlying ethical issues involved in selecting, preserving, and making available for access the private records of failed companies. In other projects the DCA is currently working on access protocols and mechanisms to assist in mitigating the risk of individual private data being improperly accessed and used.

As the project moves forward in developing and implementing a workflow for the capture and transfer of digital records to an external repository issues of maintaining the authenticity and reliability of records will be confronted. Previous efforts of NDIIPP [5][2] and other digital preservation efforts have provided many tools and technical solutions to potentially carry out the capture and transfer process. In the next stages project staff will be determining which of these tools can be implemented and operated in collaboration with Sherwood Partners.

4. CONCLUSION

Producing solutions to these challenges and successfully implementing a system to select, manage, and preserve the records of failed businesses would not only meet the objectives of the project, but such an outcome could also function as a model of methods and systems to be utilized in contexts with similar issues related to the selection and preservation of the records of private organizations. In this way the project also functions as a demonstration of the potential to preserve certain types of "at risk" records and assists in describing the process required in negotiating issues at the border between public and private digital records.

The partnership between Sherwood Partners and the DCA provides a unique opportunity to collect and preserve the "abandoned" records of failed companies that would otherwise be destroyed or unavailable for scholarly research. The collaboration between a private entity and a cultural heritage institution is itself significant and illustrates the potential of such mutually beneficial relationships to increase the preservation of important historical records. This partnership is a key element in developing a process to continually capture the records of failed companies. By seeking to increase understanding of context for actively capturing business records we hope to contribute to digital preservation research by illustrating the importance of interactions between people, organizations, and data in the laying the foundation for the processes that will collect and preserve digital records in the future.

5. REFERENCES

[1] Besek, J.M. *Copyright issues relevant to the creation of a digital archive.* Council on Library and Information Resources, Washington D.C., 2003.

[2] Boyko, A., Kunze, J., Littman, J., Madden, L., Vargas, B. *The BagIt File Packaging Format,* 2009. Retrieved June 30, 2010 from https://confluence.ucop.edu/display/Curation/BagIt

[3] Consultative Committee for Space Data Systems. Reference Model for an Open Archival Information System, 2002. Retrieved June 30, 2010 from http://public.ccsds.org/publications/archive/650x0b1.pdf

[4] Kirsch, D. "The record of business and the future of business history: Establishing a public interest in private records", *Library Trends* 57(3): 352-370, 2009.

[5] Shirky, C. Library of Congress Archive Ingest and Handling (AIHT) Final Report, 2005. Retrieved June 30, 2010 from http://www.digitalpreservation.gov/partners/aiht/high/ndiipp_aiht_final_report.pdf

TOWARDS INTEROPERABLE PRESERVATION REPOSITORIES (TIPR): THE INTER-REPOSITORY SERVICE AGREEMENT

Priscilla Caplan
Florida Center for Library Automation
Gainesville, Florida
USA

William Kehoe
Cornell University Library
Ithaca, New York
USA

Joseph Pawletko
New York University
Bobst Library
New York, New York, USA

ABSTRACT

The TIPR Project (Towards Interoperable Preservation Repositories) runs from October 2008 through September 2010. The aim of the project is to develop, test, and promote a standard format for exchanging information packages among OAIS-based repositories. This paper reviews the use cases for the transfer of information from one repository to another, reviews the Repository eXchange Format (RXP) developed by TIPR, and discusses the need for additional information not contained in the exchange package itself. It looks at two existing specifications, the Producer-Archive Interface Methodology and Into the Archive (Wege ins Archiv), in the context of inter-repository transfer. Finally it outlines information required in an inter-repository service agreement.

1. INTRODUCTION

The TIPR Project (Towards Interoperable Preservation Repositories) was begun in October 2008 with the aim of developing, testing, and promoting a standard format for exchanging information packages among OAIS-based preservation repositories. The project was premised on the idea that there are at least three real-world use cases requiring one repository to transfer an archived AIP for ingest into a different repository system:

- diversification (the owners of valuable content want it stored in multiple, heterogeneous repositories)
- succession (the source repository is ceasing operations and transferring its content to one or more other repositories)
- system migration (the repository is replacing its applications software and must migrate its archived content to the new system)

Over the past two years, the project participants have drafted and tested a package format, the Repository Exchange Package (RXP), designed to facilitate the transfer of an AIP from one repository to another. Based on the METS and PREMIS standards, the RXP describes the provenance and structure of one or more versions of a digital object.

Our prior experiences using METS and PREMIS influenced us to adopt a design philosophy for the RXP that favors constraint over flexibility. We had found that local and optional metadata elements often hinder interoperability by making exchange more difficult, impeding semantic understanding, and/or rendering the data less useful in the target systems. However, in the real world repositories are based on different software applications and run by different institutions, and there is little consistency in data models or metadata.

The TIPR approach to this dilemma is to constrain the METS and PREMIS elements in the RXP and, at the same time, to complement that constraint with some allowable flexibility, embodied in an inter-repository service agreement. The agreement complements the RXP by expressing each organization's intentions and responsibilities. The RXP bears the constrained metadata for machine transfer, while the inter-repository service agreement makes local conditions explicit, and can vary according to the circumstances and use case for any given transfer. As such, the inter-repository service agreement can be seen as a form of submission agreement between a producer and an archive.

In the next section we review the structure and content of the RXP. Section 3 reviews two specifications for the transfer of information to a digital preservation repository. In section 4 we explore the applicability of these specifications to the case of inter-repository transfer. Section 5 looks at the information required in an inter-repository service agreement.

2. A BRIEF LOOK AT THE REPOSITORY EXCHANGE PACKAGE

Conceptually, the RXP consists of three sets of files: 1) the component files of the digital object(s) being transferred; 2) metadata files describing the structure and provenance of these files; and 3) metadata files describing the structure and provenance of the package itself. Structure is described in METS documents, provenance is encoded in files containing PREMIS elements, and the digital object component files are bundled in a flat directory, their original relationships described in the METS document.

More than one version of a digital object can be packaged in an RXP, each version with its own set of structural and provenance descriptor files. These versions correspond to "representations" in PREMIS terminology.

The RXP is shown schematically in Figure 1, and is described in more detail in [1, 2, 3].

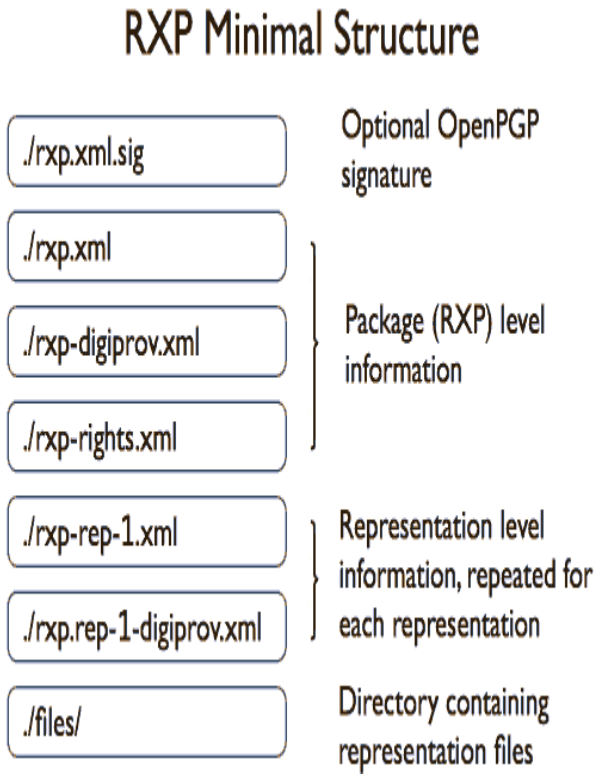


Figure 1. The structure of a Repository Exchange Package (RXP).

3. PRODUCER-ARCHIVE AGREEMENTS

It is well-accepted that the submission of content to a repository for archiving should be governed by a submission agreement. Submission agreements are addressed in the Producer-Archive Interface Methodology Abstract Standard (PAIMAS) [4] and in Into the Archive: A Guide for the information transfer to a digital repository [5].

3.1. PAIMAS

The Producer-Archive Interface Methodology is an ISO standard that builds upon the Reference Model for an Open Archival Information System [reference] and uses terms as defined in that document. Specifically, it elaborates all of the actions and negotiations that a content producer (Producer) and a repository (Archive) must take from their initial contact, through the transmission of SIPS to a repository, to the receipt and validation of the SIPs by the repository. PAIMAS is structured around phases, which must take place in order. A 46-step preliminary phase and a 36-step formal definition phase culminate in the drafting of a mutually acceptable Submission Agreement, after which shorter transfer and validation phases complete the Producer-Archive project.

While PAIMAS specifies in detail a methodology for achieving a Submission Agreement, the actual content of the agreement is largely left to be inferred from the steps leading to its creation. The Submission Agreement is described at a high level as defining the information to be transferred, the transfer process, how SIPs will be validated by the Archive, a schedule for submission, and conditions for changing or breaking the Agreement. Reporting requirements are not listed explicitly, but are implicit in transfer and validation specifications.

3.2. Into the Archive

Into the Archive (Wege ins Archiv, hereafter referred to as the "nestor Guide") is a guide produced by Germany's nestor working group on long-term preservation standards. Its aim is similar to that of PAIMAS, but it is shorter and simpler, and of a more practical than theoretical orientation. Like PAIMAS, the nestor Guide stipulates that the producer and the archive draw up a binding "ingest agreement." Ingest is defined as ending at the point where the archive has received, validated and accepted responsibility for the package, so the scope of the ingest agreement is formally the same as that of the PAIMAS submission agreement.

The nestor Guide is organized around objects, processes, and management, listing practical objectives in these areas and procedures for achieving them. Within this framework, the ingest agreement is simply another objective, rather than the end result of a long process. The ingest agreement covers much the same topics as the Submission Agreement, except that it does not include conditions for modification or termination. It goes beyond the Submission Agreement, however, in including some stipulations about how data is to be treated by the receiving archive. It requires a definition of the significant properties of the objects to be archived, the "technical environment" required for

archiving them, and agreed-upon preservation treatment ("migration agreements"). Reporting requirements are not listed as included in the ingest agreement, but reporting is a separate requirement of the information transfer process.

4. REPOSITORY TO REPOSITORY TRANSFER

The case of transfer of an AIP from one repository system to another can be seen as a special case of transfer from producer to archive. It does, however, introduce another set of contextual circumstances and some unique requirements.

4.1. Role of Producer

In both PAIMAS and the nestor Guide, the Producer is formally defined according to OAIS as the party transferring objects to the preservation repository. Both specifications are clear that the Producer does not have to be the original content creator or owner. PAIMAS explicitly allows for a third party to assume the role of Producer when there is no relationship between the Archive and the true Producer(s), giving the example of a library department entrusted with archiving a collection of CD-ROMS from a number of non-cooperating publishers. Accordingly, in the case of one repository transferring AIPs to a second repository, the sending repository could be considered a proxy producer.

Both specifications, however, carry the implicit assumption that the Producer-Archive relationship is bilateral. In a TIPR-type transfer, the relationship is more likely to be trilateral, although the alignment of players depends on the use case. In the case of diversification, the original producer (the depositor of the AIP held by the sending repository) and the proxy producer (the sending repository) are likely to be equal partners, both communicating with the archive (receiving repository). The case of succession planning may parallel case of diversification, with the original producers playing an active role, or the terminating repository may conduct all negotiations on their behalf. This case is particularly interesting as ingest concludes and in the post-ingest phase, as the receiving repository may now need to maintain relationships with a multiplicity of original producers instead of the single proxy producer (especially when the terminating repository actually ceases to exist).

The case of system migration has parallels to the succession scenario. The sending repository ceases to exist as a repository application, and the receiving repository application takes over the relationship and communications with the original producers. The institutional management of the two repository applications does not change, and of course is the same for each.

4.2. Selection of Archive

PAIMAS posits a protracted period of information exchange between Producer and Archive, at the end of which each side assesses whether or not it is desirable to continue with the project and draft a Submission Agreement. The nestor Guide assumes the two parties have already been determined and information is exchanged only to ensure the appropriate treatment of materials. In a TIPR-type transfer, the different use cases have quite different implications for the selection of a receiving repository. In the case of system migration, the organizational management of the Producer (old repository) and Archive (new repository) can be assumed to be the same, obviating the need for many PAIMAS activities. In the case of succession, the Producer (terminating repository) may not be in a position to undertake many of the steps. Only in the case of diversification are most of the PAIMAS activities likely to apply.

4.3. Selection of Content

In PAIMAS, the selection of content to be preserved is a joint responsibility of the Producer and the Archive to be worked out in the preliminary phase, although the Producer initiates the process by describing the type of information it wants to preserve. In the nestor Guide, the final selection of content falls to the archive. The assumed context is traditionally archival, where a government agency or institution exposes its entire collection to the repository, which has a legal or contractual mandate to assume responsibility for items which meet certain criteria.

In a TIPR-type transfer, the three use cases have different implications for selection. In the case of diversification, it is almost certainly the original or proxy Producer who will identify specific content and seek a repository most capable of preserving it. In the case of system migration, there is likely to be no selection at all, the assumption being that all of the content in the old system will be transferred to the new. In the case of succession, either all of the terminating repository's content will be transferred to a single receiving repository, or variously defined subsets of content (for example, by media type, or by original owner) will be identified for transfer to different repositories. While the receiving repository will have some say in what it will agree to take, in no case does it have primary responsibility for selection. In this respect PAIMAS models selection better than the nestor Guide.

4.4. SIP Creation

Both PAIMAS and the nestor Guide assume the Producer is creating an original SIP (i.e., a SIP for first-time archiving). In the case of repository to repository transfer of a SIP created from a previously archived

AIP, the sending repository has additional constraints; for example, it may not be able to obtain additional metadata from the original producer. At the same time, the sending repository is likely to have enriched the original AIP with metadata of its own, such as format-specific details, validation results, and processing history. While these factors will complicate the negotiation of a transfer project, the existence of a standard transfer format such as the RXP dramatically simplifies and/or obviates the need for a number of steps defined in PAIMAS.

4.5. Role of Agreement

In the nestor Guide, the ingest agreement is a single objective covering only the specifics of ingest, although the other objectives and procedures in the guide go well beyond those needed for ingest to the subsequent preservation treatment, access control, and rights management of objects. PAIMAS similarly describes a fairly restricted Submission Agreement, but includes consideration of future financial, technical and management issues in the steps leading up to the Agreement. In fact, although both specifications profess their scope is the transfer of information, the transfer and ingest of SIPs can not realistically be considered outside of the broader context of a long-term archiving agreement.

An inter-repository service agreement, as envisioned by TIPR, must clarify the technical details of a specific act of transfer, but it must also explicitly address post-ingest preservation treatment, ongoing access controls, rights, and communications.

5. THE INTER-REPOSITORY SERVICE AGREEMENT

The last section explored the general applicability of PAIMAS and the nestor Guide to the case of repository to repository transfer. This section focuses specifically on the inter-repository service agreement as a variant of the Submission or ingest agreement. The TIPR approach was to define a relatively rigid transfer format for machine processing and rely on the inter-repository service agreement to provide context, meaning, and external stipulations.

5.1. Meaning of RXP Elements

The RXP defines a standard place to put some critical pieces of information, but does not define code lists (controlled vocabulary) or semantics for the content. For example, the sending repository is identified in the *agent* element of the METS header in rxp.xml. The value used for identification must be negotiated between the parties and documented in the inter-repository service agreement. The receiving repository may need to predefine an agent record, add a mapping to a processing table, etc. This also applies to identification of the original producer and the original rights holder.

5.2. Transfer Details

The RXP specification defines only a transfer format, and leaves details of the transfer protocol to be determined by the parties. In the TIPR project, test packages were bundled according to the BagIt specification and transmitted via HTTP, but they could equally as well have been zipped in native form and shipped on a portable drive. The inter-repository service agreement should document agreement on the transfer mechanism and serialization, and manifests used (if any). In addition, communication between repositories and the handling of transmission errors must be specified. Transfer requirements are well covered in PAIMAS and the nestor Guide.

5.3. Actions to be taken on Ingest

Actions taken by the receiving repository after successful transfer are out of scope for TIPR and the RXP. Whether and how the receiving repository performs quarantine, validates packages and files, gives notification of rejection or successful ingest, and gives notification of anomalies and non-fatal errors all must be agreed upon and documented. Although much of this is covered in PAIMAS and the nestor Guide, both specifications stop at the point where the receiving repository has validated and accepted responsibility for the SIPs, which for some preservation repository systems may be far in advance of the creation and storage of a new AIP.

A complication in repository-to-repository transfer is the circumstance that in some cases notification should be made to the sending repository, and in other cases to the original owner of the content. Especially in the case of succession, the receiving repository may need to establish an ongoing relationship with the original owner(s).

5.4. Archiving Policies and Responsibilities of the Receiving Repository

Repository systems differ greatly in their internal data models and the type and amount of metadata they store. The TIPR project asserts that preservation repositories engaging in package exchange should be capable of understanding METS stucture and the semantics of PREMIS events. Beyond that, what metadata will be retained and what will be understood (in the sense that it will be maintained in a usable fashion) by the receiving repository is a matter for negotiation and documentation. Similarly preservation treatment, retention of versions, ongoing reporting, future dissemination and access are all appropriate for documentation in the inter-repository service agreement.

5.5. Rights and Permissions

The TIPR RXP provides a place to record package-level rights. TIPR partners assumed repositories would use PREMIS rights statements, but any XML-encoded rights schema could be used if agreed-upon and included in the inter-repository service agreement. Rights governing individual files in the package, whether metadata or content, is not covered by the RXP specification and is entirely a mater of agreement among transfer partners.

5.6. Financial Arrangements

Costs involved in the transfer project and ongoing custodial costs should both be documented along with the method for identifying and billing the appropriate party. In the case of succession, a likely scenario is that fixed costs of the transfer project are assumed by the terminating repository but ongoing custodial costs must be charged to the original producers.

5.7. Legal issues

The source repository can be assumed to have a standing legal agreement with its own Producers clarifying intellectual property rights, responsibility for copyright infringement, and liabilities and warranties governing damage to content, treatment of content, and provision of services. In the case of repository-to-repository transfer, the legal relationship between the Producers and the original repository may carry over to the receiving repository but is more likely to require re-negotiation. Legal issues pertaining to the source repository must be considered separately from those pertaining to the original depositors, and documented in the inter-repository service agreement.

6. CONCLUSION

Two existing standards address the transfer of information from a producer (in OAIS terms) to a preservation repository. Although neither explicitly restrict their applicability to the original producer or content owner, neither consider the special case of a repository to repository transfer. The three use cases of interest to the TIPR project have different implications for the methodology of transfer and the circumstances considered. An inter-repository service agreement has much in common with a Submission (ingest) agreement, but must have a longer-term scope and take into account two producers, the producers of the original SIP and the proxy producer, the repository that creates the RXP for transfer.

7. REFERENCES

[1] Caplan, P., "Repository to Repository Transfer of Enriched Archival Information Packages", in *D-Lib Magazine*, v.14 no.11/12, 2008. Available at http://www.dlib.org/dlib/november08/caplan/11caplan.html

[2] Caplan, P., "Towards Interoperable Preservation Repositories (TIPR)", U.S. Workshop on Roadmap for Digital Preservation Interoperability Framework, 2010. Available at http://ddp.nist.gov/workshop/papers/03_08_Caplan_TIPR.pdf

[3] Caplan, P., Kehoe, W., Pawletko, J., "Towards Interoperable Preservation Repositories", *International Journal of Digital Curation*, v.5 no.1, 2010.

[4] Consultative Committee for Space Data Systems, Producer-Archive Interface Abstract Standard (CCSDS 651.0-B-1 Blue Book (2004). Available at http://public.ccsds.org/publications/archive/651x0b1.pdf

[5] nestor working group for long-term preservation standards 2009 – Into the Archive – a guide for the information transfer to a digital repository. Draft for public comment. Available at http://files.d-nb.de/nestor/materialien/nestor_mat_10_en.pdf

Tutorials

THE NEXT-GENERATION JHOVE2 FRAMEWORK AND APPLICATION

Stephen Abrams
California Digital Library
Oakland, CA 94612, US

Tom Cramer
Stanford University
Stanford, CA 94305, US

Sheila Morrissey
Portico
Princeton, NJ 08450, US

ABSTRACT

JHOVE2 is a Java framework and application for next-generation format-aware characterization of digital objects [1]. Characterization is the process of deriving representation information about a formatted digital object that is indicative of its significant nature and useful for purposes of classification, analysis, and use in digital curation, preservation, and repository contexts. JHOVE2 supports four specific aspects of characterization: (1) identification, the determination of the presumptive format of a digital object on the basis of suggestive extrinsic hints and intrinsic signatures; (2) validation, the determination of the level of conformance to the normative syntactic and semantic rules of the object's format; (3) feature extraction, the process of reporting the intrinsic properties of an object significant for purposes of classification, analysis, and use; and (4) assessment, the determination of the level of acceptability of an object for a specific purpose on the basis of locally-defined policy rules.

The object of JHOVE2 characterization can be a file, a subset of a file, or an aggregation of an arbitrary number of files that collectively represent a single coherent digital object. JHOVE2 can automatically process objects that are arbitrarily nested in containers, such as file system directories or Zip files.

The JHOVE2 project is a collaborative undertaking of the California Digital Library, Portico, and Stanford University, with generous funding from the Library of Congress. Additional information about JHOVE2 can be found on the project wiki [2]. The project seeks to build on the success of the original JHOVE characterization tool [3] by addressing known limitations and offering significant new functions, including: streamlined APIs with increased modularization, uniform design patterns, and comprehensive documentation; object-focused, rather than file-focused, characterization; signature-based file-level identification using DROID [4]; aggregate-level identification based on configurable file system naming conventions; rules-based; extensive user configuration of plug-in modules, characterization strategies, and formatted results using the Spring dependency injection framework [5]; and performance improvements using Java buffered I/O (java.nio).

The main topics covered during the tutorial are: the role of characterization in digital curation and preservation workflows; an overview of the JHOVE2 project: requirements, methodology, and deliverables; demonstration of the JHOVE2 application; architectural review of the JHOVE2 framework and Java APIs; integration of JHOVE2 technology into existing or planned systems, services, and workflows; third-party development of conformant JHOVE2 modules; and building and sustaining the JHOVE2 user community. JHOVE2 is made freely available under the terms of the BSD open source license.

This tutorial is an updated and expanded version of the workshop presented at iPRES 2009 in San Francisco [6]. This tutorial will closely follow the production release of JHOVE2 and will incorporate significant new material arising from the second year of project work. The targeted audience for the tutorial includes digital curation, preservation, and repository managers, analysts, tool users and developers, and other practitioners and technologists whose work is dependent on an understanding of the format and pertinent characteristics of digital assets.

1. REFERENCES

[1] Abrams, S., Morrissey, S., and Cramer, T. "What? So what?": The next-generation JHOVE2 architecture for format-aware characterization. International Journal of Digital Curation 4, 3 (2009), 123-136.

[2] California Digital Library. JHOVE2 Home. http://jhove2.org/

[3] Harvard University Library. JHOVE – JSTOR/Harvard Object Validation Environment. http://hul.harvard.edu/jhove

[4] The National Archives, DROID. http://sourceforge.net/projects/droid

[5] SpringSource. The Standard for Enterprise Java Development. http://www.springsource.com/products/enterprise

PREMIS TUTORIAL: AN EXPLORATION OF THE PREMIS DATA DICTIONARY FOR PRESERVATION

PREMIS Editorial Committee
Submitted by Rebecca Guenther
Chair
Library of Congress
Washington, DC

ABSTRACT

The PREMIS Data Dictionary for Preservation Metadata is a specification that provides a key piece of infrastructure for digital preservation activities, playing a vital role in enabling the effective management, discovery, and re-usability of digital information. Preservation metadata provides provenance information, documents preservation activity, identifies technical features, and aids in verifying the authenticity of digital objects. PREMIS is a core set of metadata elements (called "semantic units") recommended for use in all preservation repositories regardless of the type of materials archived, the type of institution, and the preservation strategies employed. This tutorial provides an introduction to PREMIS and its data model and an examination of the semantic units in the Data Dictionary organized by the entities in the PREMIS data model, objects, events, agents and rights. In addition it presents examples of PREMIS metadata and a discussion of implementation considerations, particularly using PREMIS in XML and with the Metadata Encoding and Transmission Standard (METS). It will include examples of implementation experiences.

The PREMIS Data Dictionary was originally developed by the Preservation Metadata: Implementation Strategies (PREMIS) Working Group in 2005 and revised in 2008. It is maintained by the PREMIS Editorial Committee and the PREMIS Maintenance Activity is managed by the Library of Congress.

The tutorial aims at developing and spreading awareness and knowledge about metadata to support the long term preservation of digital objects. The tutorial will benefit individuals and institutions interested in implementing PREMIS metadata for the long-term management and preservation of their digital information but who have limited experience in implementation. Potential audience includes cultural heritage operators, researchers and technology developers, professional educators, and others involved in management and preservation of digital resources.

PRESENTERS

Priscilla Caplan: Priscilla Caplan is Assistant Director for Digital Library Services at the Florida Center for Library Automation, where she oversees the Florida Digital Archive, a digital preservation repository for the use of the public universities of Florida.

Angela Dappert: Angela Dappert is a Senior Analyst at the British Library. Her current focus is on conceptual modeling of preservation planning and characterization within the Planets project.

Markus Enders: Markus Enders is a Technical Architect at the British Library and has contributed to several METS and PREMIS based digital library projects developing METS profiles for eJournal and newspaper preservation.

Karin Bredenberg: Karin Bredenberg is a Programmer at the Swedish National Archives where she works with Swedish adaptations of international archival metadata standards.

TUTORIAL: LOGICAL AND BIT-STREAM PRESERVATION INTEGRATED. DIGITAL PRESERVATION USING PLATO AND EPRINTS

Hannes Kulovits **Andreas Rauber**
Vienna University of Technology
Institute of Software Technology and Interactive Systems
Austria

David Tarrant **Steve Hitchcock**
University of Southampton
Electronics and Computer Science
UK

ABSTRACT

The rapid technological changes in today's information landscape have considerably turned the preservation of digital information into a pressing challenge. The aim of an institutional repository has evolved in the last decade from the simple need to provide material with a persistent online home, to an infrastructure that facilitates services on complex collections of digital objects.

Digital librarians have long acknowledged the preservation function as a vital back office service that is central to the role of repository. However, preservation is often sidelined due to the practical constraints of running a repository. Dealing with institutional-scale ingests and quality assurance with minimal staff and investment rarely leaves sufficient capacity for engaging with a preservation agenda. A lot of different strategies, i.e. preservation actions, have been proposed to tackle this challenge: migration and emulation are the most prominent ones. However, which strategy to choose, and subsequently which tools to select to implement it, poses significant challenges. The creation of a concrete plan for preserving an institution's collection of digital objects requires the evaluation of possible preservation solutions against clearly defined and measurable criteria.

This tutorial shows attendees the latest facilities in the EPrints[1] open source repository platform for dealing with preservation tasks in a practical and achievable way, and new mechanisms for integrating the repository with the cloud and the user desktop, in order to be able to offer a trusted and managed storage solution to end users.

Furthermore, attendees will create a preservation plan on the basis of a representative scenario and receive an accountable and informed recommendation for a particular preservation action. The whole preservation planning process will be supported by Plato[2], a decision support tool that implements a solid preservation planning approach and integrates services for content characterisation, preservation action and automatic object comparison to provide maximum support for preservation planning endeavours. Attendees will then enact the preservation plan created in Plato by uploading it to the EPrints repository. By uploading the preservation plan EPrints automatically carries out the recommended preservation action, e.g. migrating all GIF images in a repository to PNG, and links the plan to both the original and the migrated file.

The benefit of this tutorial is the grounding of digital curation advice and theory into achievable good practice that delivers helpful services to end users for their familiar personal desktop environments and new cloud services.

PRESENTERS

David Tarrant has been central in the development of the EPrints storage and preservation infrastructure through his involvement in the JISC Preserv project (http://preserv.eprints.org). He will be presenting on the day and providing technical infrastructure support for the practical exercises.

Steve Hitchcock has been involved with Institional Repositories and Open Access from their outset, helping launch EPrints in 2001 as the first OAI-compliant repository software. Steve has been working with institutions using EPrints and encouraging them to become participating members of the Community. Steve has also been actively involved at the core of all the EPrints preservation projects dating back to 2006 with Preserv1.

Andreas Rauber is Associate Professor at the Department of Software Technology and Interactive Systems at the Vienna University of Technology. He is actively involved in several research projects in the field of Digital Libraries, focusing on the organization and exploration of large information spaces, as well as Web archiving and digital preservation. His research interests cover the broad scope of digital libraries, including specifically text and music information retrieval and organization, information visualization, as well as data analysis and neural computation. He is involved in numerous initiatives in the area of digital preservation, such as DPE - Digital Preservation Europe; Planets - Preservation and Long-term Access Networked Services; nestor - Network of expertise in Digital long-term preservation. He has been lecturing extensively on this

[1] http://www.eprints.org/
[2] http://www.ifs.tuwien.ac.at/dp/plato

subject at different universities, as part of the DELOS and nestor summer schools on digital reservation, as well as during a range of training events on digital preservation.

Hannes Kulovits is currently a researcher at the Department of Software Technology and Interactive Systems at the Vienna University of Technology. He received his Master in Business Informatics from the Vienna University of Technology in 2005. He is actively involved in several research projects in the field of Digital Preservation where his main focus lies in Preservation Planning and Recommender Systems.

TUTORIAL: PERSONAL DIGITAL ARCHIVING

Ellyssa Kroski
Information Services Technologist
Barnard College Library
New York, NY USA

ABSTRACT

Today more and more of our lives are becoming digital. Everything from family photographs, music files, video footage, and correspondence to medical records, bookmarks, documents, and even ideas are now available in electronic form. This makes access quick & convenient, but how do we save all of these digital assets for the long term? Most of us have experienced personal data loss at one time or another due to hard drive failure, file corruption, technology obsolescence, or accidental file deletion. What should we be doing right now to safeguard our digital creations? This hands-on session will explain the process of creating and executing an action plan for archiving personal digital assets, deciding what to store, consolidating multiple file versions, and cataloguing resources. This workshop will explore both local storage media and cloud services as well as institutional & disciplinary repositories. Learn to plan & execute the archiving of your own personal digital assets as well as how to teach your patrons to do this for themselves.

TUTORIAL: STABILITY OF DIGITAL RESOURCES ON THE INTERNET AND STRATEGIES OF PERSISTENT IDENTIFIERS

Jürgen Kett
Deutsche Nationalbibliothek
Digital Services/IT

Maurizio Lunghi
Fondazione Rinascimento Digitale
Firenze, Italy

ABSTRACT

Within research activity worldwide about digital preservation many studies, criteria sets, tools, strategies, standards and best practices have been developed by the practitioners: one of these technology families is the Persistent Identifier (PI) to grant stability of digital objects over time. PIs give things that we use or talk about in information systems a unique and stable name. While the location of a resource may change, its PI remains the same. Persistent identification of Internet resources is a crucial issue for almost all the sectors of the future information society. In particular, in the cultural/scientific digital library applications, the instability of URLs reduces the credibility of digital resources which is a serious drawback especially for researchers.

There are various concepts and schemes for persistent identification that pretend to solve this problem: Digital Object Identifier (DOI) [7], Persistent Uniform Resource Locator (PURL) [5], Archival Resource Key (ARK) [2] and Uniform Resource Name (URN) [6] to name a few. They all share common goals but there are indeed important differences between these approaches with respect to the use cases, communities and business models towards they are directed. Recently the diversity of possible solutions is getting even more confusing: The PI systems mentioned above all primarily focus on the identification of web resources that are meant to be available in the long term and are subject to long-term preservation. But with the raise of the Data Web, which is driven by the success of social networks and the Linking Open Data movement, the identification of non-digital entities (like real-world objects, events, places and persons) and abstract concepts is getting more and more important. Especially in this context the traditional PI systems compete with lightweight solutions like "Cool URIs" [3] and Hashtags.

But the key qualities of a PI service are mostly independent of the scheme it uses. They concern trust and reliability. No technology can grant a level of service in any case without a trustable organisation and clear defined policies: it is well known that digital preservation is more an organisational issue than a technical one. European activities, like the development of the Europeana Resolution Discovery Service [6] and PersID [1], focus on the harmonization of the national PI strategies and embed all these existing approaches into a shared infrastructure. The aim is to establish a transparent and trusted service for the cultural and academic sector. The crucial question is: How much und what kind of regulation by public authorities does the web of culture and research need?

In this tutorial we explain the importance of trusted Persistent Identifier services for the web's evolution and present a survey of available technologies and current practices. The tutorial starts with introduction of the problems PI systems try to solve today and those that they will have to address in the future. Then we will present a survey of available technologies and the major initiatives world wide, talk about their commonalities and differences and highlight the most important issues and problems with the current situation. More in detail the Europeana Resolution Discovery Service (ERDS) and the PersID goals and plans will be outlined. The tutorial will close with an open debate or round table on "use cases and user requirements for a PI system".

The tutorial is directed towards people in charge of digital repositories, institutions working in the context of linked data, authors of digital contents, software companies developing archival solutions and digital library applications, researchers and students working on digital libraries for cultural and scientific resources.

REFERENCES

[1] About PersID. www.surffoundation.nl/wiki/display/persid/About

[2] ARK: Archival Resource Key. www.cdlib.org/inside/diglib/ark/

[3] Cool URIs for the Semantic Web. www.w3.org/TR/cooluris/

[4] Europeana Connect: Results and Resources. www.europeanaconnect.eu/results-and-resources.php

[5] PURL Homepage. www.purl.org

[6] Uniform Resource Names (urn) – Charter. www.datatracker.ietf.org/wg/urn/charter/

[7] Welcome to the DOI® System. www.doi.org

Author Index