iPRES 2014

MELBOURNE | 6-10 OCTOBER

Proceedings of the 11th International Conference on Digital Preservation

Proceedings of the 11th International Conference on Digital Preservation

ISBN 978-0-642-27881-4

Layout and compilation: Anna Metcalfe, State Library of Victoria

State Library of Victoria
328 Swanston Street
Melbourne, Victoria 3000
slv.vic.gov.au

Presented by

State Library of Victoria

NATIONAL LIBRARY OF AUSTRALIA

State Government Victoria

Sponsored by

Microsoft

ExLibris

Preservica Digital Preservation

EMC²

CAARA Council of Australasian Archives and Records Authorities

CITY OF MELBOURNE

iPRES 2014

MELBOURNE | 6-10 OCTOBER

Proceedings of the 11th International Conference on Digital Preservation

Edited by Serena Coates, Ross King, Steve Knight, Christopher (Cal) Lee, Peter McKinney, Erin O'Meara, and David Pearson

Conference Committee

Co-Convenors

Sue Roberts	State Library of Victoria, Australia
Anne-Marie Schwirtlich	National Library of Australia, Australia

General Chairs

Liz Jesty	State Library of Victoria, Australia
David Pearson	National Library of Australia, Australia

Programme Chairs

Steve Knight	National Library of New Zealand, New Zealand
Christopher (Cal) Lee	University of North Carolina at Chapel Hill, United States

Workshop/Panels/Tutorials Chairs

Peter McKinney	National Library of New Zealand, New Zealand
Erin O'Meara	Gates Archive, United States

Poster Chairs

Serena Coates	State Library of Queensland, Australia
Ross King	Austrian Institute of Technology, Austria

Local Organising Committee

Emma Bloom	State Library of Victoria, Australia
Annette Brown	State Library of Victoria, Australia
Phillip Clifford	State Library of Victoria, Australia
Justine Heazlewood	Public Records Office of Victoria, Australia
Liz Jesty	State Library of Victoria, Australia
Teula Morgan	Swinburne University, Australia
David Pearson	National Library of Australia, Australia
Sue Roberts	State Library of Victoria, Australia
Gail Schmidt	State Library of Victoria, Australia
Sarah Slade	State Library of Victoria, Australia
Janice Van de Velde	State Library of Victoria, Australia
Nerida Webster	State Library of Victoria, Australia

Programme Committee

Stephen Abrams	California Digital Library, United States
Reinhard Altenhöner	German National Library, Germany
David Anderson	University of Portsmouth, United Kingdom
José Borbinha	Lisbon Technical University, Portugal
Raju Buddharaju	National Library Board Singapore, Singapore
Thu-Mai Christian	University of North Carolina, United States
Euan Cochrane	Yale University, United States
Angela Dappert	DPC, United Kingdom
Kevin DeVorsey	National Archives and Records Administration, United States
Wendy Duff	University of Toronto, Canada
Kresimir Duratec	TU Wien, Austria
Michael Factor	IBM Haifa, Israel
Ed Fay	LSE, United Kingdom
Andrea Goethals	Harvard University, United States
Neil Grindley	JISC, United Kingdom
Mariella Guercio	University of Roma Sapienza, Italy
Ross Harvey	RMIT, Australia
Andrew Jackson	British Library, United Kingdom
Leslie Johnston	Library of Congress, United States
Michelle Lindlar	German National Library of Science and Technology, Germany
Hansueli Locher	National Library of Switzerland, Switzerland
Jens Ludwig Göttingen	State and University Library, Germany
Nancy McGovern	MIT, United States
Don Mennerich	New York Public Library, United States
Courtney Mumma	Artefactual Systems, Canada
Heike Neuroth	Göttingen State and University Library, Germany
Tomasz Parkola	Poznan Supercomputing and Networking Center, Poland
Natalie Pang	Nanyang Technological University, Singapore
Andreas Rauber	Vienna University of Technology, Austria
Thomas Risse	L3S, Germany
Seamus Ross	University of Toronto, Canada
Sven Schlarb	ONB, Austria
Daisy Selematsela	National Research Foundation, South Africa
Robert Sharpe	Tessella, United Kingdom
Barbara Sierman	National Library of the Netherlands, Netherlands
Shigeo Sugimoto	University of Tsukuba, Japan
Kara Van Malssen	AV Preserve, United States
Lucia Maria Velloso de Oliveira	Fundacao Casa de Rui Barbosa, Brazil
Paul Wheatley	Paul Wheatley Consulting Ltd, United Kingdom
Andrew Wilson	Queensland State Archives, Australia
Qu Yunpeng	National Library of China, China
Zhixiong Zhang	National Science Library, China
Eld Zierau	Royal Library of Denmark, Denmark

PREFACE

Co-convenors Anne-Marie Schwirtlich (Director-General of the National Library of Australia) and Sue Roberts (CEO and State Librarian of Victoria) were delighted to be able to welcome 200 delegates (115 from Australia and 85 from 17 other countries) to Melbourne for the 11th International Conference on Digital Preservation (iPres), held from 6–10 October 2014.

The conference was structured around a programme of workshops on the Monday and Tuesday, papers, posters and panels during the core conference on Wednesday and Thursday, a plenary panel and closing remarks on Friday morning, and finished up with more workshops on Friday afternoon.

We received 92 total submissions and accepted 69 (22 full papers, 15 short papers and 13 posters, 5 demos, 6 workshops, 5 tutorials and 3 panels). The acceptance rate for research paper submissions was 51% (18 out of 35).

Keynotes

Dr Shaun Hendy (Professor of Physics and Director of Te Pūnaha Matatini – the Centre for Complex Systems and Networks – at the University of Auckland) presented on the connections between 'Preservation, Innovation and Collaboration'. He reinforced the notion that 'we must collaborate to innovate' as digital preservation becomes increasingly important as governments and businesses increasingly move to adopt more data-driven decision-making and policy. Policy evaluation may take decades, so policy makers and researchers need rich digital records of decision-making processes and outputs to inform policy evaluation in multiple sectors including research, education and innovation.

Dr Ross Wilkinson (Executive Director of the Australian National Data Service) opened Tuesday's session with a presentation on 'The value of digital preservation: Exploring the benefits of preserved data to researchers, institutions and nations'. He discussed the different perspectives of researchers, research institutions, and the public at large on the value of data. He noted that there is a variety of interests that need to be taken into account when considering the preservation of data including the researchers who create data, the government and taxpayers who frequently fund research as well as research and collecting institutions that are often responsible for the long-term safekeeping of research outputs.

Dr Herbert Van de Sompel (Leader of the Prototyping Team at the Research Library of the Los Alamos National Laboratory) in his presentation, 'When I say NOW, it's already over', noted that the pace and extent of web-based communication is 'astounding' and brings with it a focus on an eternal Now and a risk of neglecting the Past. He then explored some of the challenges of providing appropriate access to remnants of the ephemeral web information environment of the Now at some point in the Future with a particular emphasis on the complexity of assuring the temporal coherence of embedded web resources such as images and style sheets.

The programme

The conference this year was structured around two key strands – research and innovative practice. The purpose of this distinction was to promote both academic/research work and work that is clearly rooted in the actual experience of institutions undertaking digital preservation (while acknowledging that some work encapsulates both of these strands).

We had an excellent array of papers and posters with the award for Best Paper (sponsored by Ex Libris) going to Miksa, Vieira, Barateiro, and Rauber for their paper 'VPlan – Ontology for collection of process version data'. The judges noted that 'this paper introduces the VPlan ontology for managing significant characteristics of preserved processes and workflows that can be used for the automated verification of future redeployments of those workflows. By facilitating confidence in the independent replicability of scholarly claims based on computational analyses, VPlan helps to ensure the trustworthiness and creditability of scholarly advancement'.

Honourable mentions also went to Gattuso and McKinney, 'Converting WordStar to HTML4' and Graf, Gordea, and Ryan, 'A model for format endangerment analysis using fuzzy logic'.

The award for Best Poster (sponsored by CAARA – the Council of Australasian Archives and Records Authorities) went to Bähr, Rechert, Liebetraut and Lindlar for their poster on 'Functional Access to Electronic Media Collections using Emulation-as-a-Service'.

Papers covered a wide array of preservation topics including migration and emulation, file format management, registries and linked data, funding models, education and training, personal archiving and software-based art, web archiving, metadata and persistent identifiers.

A new addition to this year's conference was the Digital Preservation Systems Showcase in which a set of vendors presented their systems' implementation of a pre-defined set of functions, thereby providing a unique opportunity to view digital preservation systems in an 'apples to apples' comparison. The systems presented in the showcase were DuraCloud, Archivematica, RODA from KEEP Solutions, Preservica and Rosetta.

The showcase divided digital preservation functionality into four large categories:
- **How do we get content in** – which included ingest flows/methods, preconditioning/pre-ingest preparation, format identification, metadata extraction, fixity checking/assignation and virus checking.
- **How do we manage and preserve the content** – which included intellectual management, risk analysis, preservation planning, preservation execution, repository management (queries, monitoring, analysis, updates) and exception handling.
- **How can the content be accessed from the system** – which included derivative generation (static, on-the-fly, options of types), access rights, complex materials, handing over to other access methods and export of data.
- **Other considerations** – which included flexibility/interoperability of the system, exit strategy, Archival Information Package creation, relationships to PREMIS and other metadata schemas, data models, provenance, testing and storage.

Acknowledgments

This year's conference was generously supported by sponsors Preservica, Ex Libris, EMC, City of Melbourne, Microsoft, and the Council of Australasian Archives and Records Authorities.

The conference banquet (sponsored by Preservica) was held in the lovely Queen's Hall at the State Library of Victoria and provided an excellent opportunity for all the delegates to mingle, network, share information and generally enjoy the opportunity to talk to colleagues from near and afar.

The Organising Committee was very pleased with the success of the conference, and wishes to acknowledge the contribution of the many members of the Programme Committee who helped ensure the high quality of the papers, posters and ancillary events attached to iPres this year. The Programme Co-chairs would also like to acknowledge the efforts of the local organisers in ensuring the smooth running of the conference and the warm welcome extended to delegates which helped create a collegial atmosphere throughout the event.

Finally, we would like to acknowledge the volunteers from the Royal Melbourne Institute of Technology (RMIT), Charles Sturt University and the National Library of Australia who undertook so much of the behind-the-scenes work that made iPres 2014 so successful. We now pass the baton on to the University of North Carolina at Chapel Hill who will be hosting iPres in 2015. We look forward to seeing you all there.

Steve Knight and Christopher (Cal) Lee
Programme Chairs, iPres 2014

Monday 6 October 2014
Workshops

8.30AM–5PM	REGISTRATION			
9AM–1PM		**Defining a Roadmap for Economically Efficient Digital Curation – A 4C Project Workshop** Neil Grindley, Katarina Haage, Paul Stokes	**Born Digital Appraisal, Ingest, and Processing** Jessica Moran (Chair), Leigh Rosin, Douglas Elford, Emma Jolley, Somaya Langley, Donald Mennerich, Ben Fino-Radin, Christopher A. Lee, Erin O'Meara	**PREMIS Implementation Fair Workshop** Peter McKinney, Eld Zierau, Rebecca Guenther
1–2PM	LUNCH VENUE: QUEEN'S HALL			
2–5PM		**ICA-AtoM, Archivematica and Digital Preservation** Lise Summers, Meg Travers	**Preserving Data to Preserving Research: Curation of Process and Context** Angela Dappert, Rudolf Mayer, Stefan Pröll, Andreas Rauber, Raul Palma, Kevin Page, Daniel Garijo	**Memento. Uniform and Robust Access to Resource Versions** Herbert Van de Sompel

Tuesday 7 October 2014

Workshops

8.30AM–5PM	REGISTRATION			
9AM–1PM	**Digital Preservation Systems Showcase**	**Modelling file formats and technical environments using the NSLA Digital Preservation Technical Registry (DPTR)** Jan Hutar, Ross Spencer, Libor Coufal, Kevin DeVorsey, Jay Gattuso, Steve Knight, Peter McKinney		**Acquiring and processing Born-digital data using the BitCurator environment** Christopher A. Lee
1–2PM	LUNCH VENUE: QUEEN'S HALL			
2–5PM	**Digital Preservation Systems Showcase** (Note: continuation of morning workshop; ends 5.15pm)	**Modelling file formats and technical environments using the NSLA Digital Preservation Technical Registry (DPTR)** (Note: continuation of morning workshop)		**Acquiring and processing Born-digital data using the BitCurator environment** (Note: continuation of morning workshop)
5.30–7.30PM	**Welcome reception:** **Sue Roberts** **CEO and State Librarian** **GENEROUSLY SPONSORED BY EMC**			

Wednesday 8 October 2014

8AM–5PM	REGISTRATION		
9–9.20AM	**Opening and welcome: Anne-Marie Schwirtlich (Co-Convenor)** **Director-General National Library of Australia**		
9.25–10.25AM	**Keynote address: Preservation, Innovation and Collaboration** **Professor Shaun Hendy FRSNZ** **MacDiarmid Institute for Advanced Materials and Nanotechnology** **Professor of Physics and Director of Te Pūnaha Matatini – the Centre for Complex Systems and Networks – at the University of Auckland** Chair: Steve Knight, National Library of New Zealand		
10.25–10.30AM	HOUSEKEEPING		
10.30–10.55AM	MORNING TEA VENUE: CONFERENCE CENTRE		
11–11.30AM	**Session Chair: Erin O'Meara, Gates Archive** **Linked Data Registry: A New Approach To Technical Registries** Maïté Braud, James Carr, Kevin Leroux, Joseph Rogers, Robert Sharpe	**Session Chair: Janet Delve, University of Portsmouth** **New Perspectives on Economic Modeling for Digital Curation** Neil Grindley, Ulla Bøgvad, Hervé L'hours	**Session Chair: David Anderson, University of Portsmouth** 11–11.20am **Achieving Canonical PDF Validation** Duff Johnson 11.20–11.40am **Making the strange familiar: Bridging boundaries on database preservation projects** Peter Francis, Alan Kong
11.30AM–12PM	**A next generation technical registry: moving practice forward** Peter McKinney, Steve Knight, Jay Gattuso, David Pearson, Libor Coufal, Kevin Devorsey, David Anderson, Janet Delve, Ross Spencer, Jan Hutař	11.30–11.50am **Developing costing-models for emulation based access in scientific libraries** Euan Cochrane, Dirk Von Suchodoletz, Klaus Rechert	11.40am–12pm **Addressing the personal digital archives needs of a contemporary artist** Sam Meister
12–12.30PM	**Automatic Discovery of Preservation Alternatives Supported by Community Maintained Knowledge Bases** Rudolf Mayer, Johannes Binder, Stephan Strodl	11.50am–12.10pm **Networked Instruction for Research Data Curation Education: The CRADLE Project** Helen Tibbo, Thu-Mai Christian	**Virtualisation as a Tool for the Conservation of Software-Based Artworks** Patricia Falcao, Alistair Ashe, Brian Jones
12.30–1.30PM	LUNCH VENUE: QUEEN'S HALL		

1.30–2PM	**Session Chair: Nancy McGovern, Massachusetts Institute of Technology** **Risk Driven Selection of Preservation Activities for Increasing Sustainability of Open Source Systems and Workflows** Tomasz Miksa, Rudolf Mayer, Stephan Strodl, Ricardo Vieira, Goncalo Antunes, Andreas Rauber	**Panel: Getting to Digital Preservation Tools that “just work”** Paul Wheatley, Stephen Abrams, David Clipsham, Janet Delve, Ed Fay, Christopher A. Lee, Andrea Goethels	**Session Chair: Serena Coates, State Library of Queensland** **Management and Orchestration of Distributed Data Sources to Simplify Access to Emulation-as-a-Service** Thomas Liebetraut, Klaus Rechert
2–2.30PM	**Epimenides: Interoperability Reasoning for Digital Preservation** Yannis Kargakis, Yannis Tzitzikas, René van Horik		**A Persistent Identifier e-Infrastructure** Barbara Bazzanella
2.30–3PM	**A Novel Metadata Standard for Multimedia Preservation** Walter Allasia, Werner Bailer, Sergiu Gordea, Wo Chang		**Access and Preservation in the cloud: Lessons from operating Preservica Cloud Edition** Kevin O’Farrelly, Alan Gairey, James Carr, Maite Braud, Robert Sharpe (Note: ends 2.50pm)
3–3.25PM	**AFTERNOON TEA** VENUE: THE COURTYARD		
3.30–4PM	**Session Chair: David Pearson, National Library of Australia** **Sustainability Assessments at the British Library: Formats, Frameworks, & Findings** Maureen Pennock, Paul Wheatley, Peter May	**Panel: Preserving Government Business Systems** Cassie Findlay, Neal Fitzgerald, Andrew Waugh, Richard Lehane	**Session Chair: Andrea Goethals, Harvard University** **Converting WordStar to HTML4** Jay Gattuso, Peter McKinney
4–4.30PM	**A Model for Format Endangerment Analysis using Fuzzy Logic** Roman Graf, Sergiu Gordea		**VPlan – Ontology for Collection of Process Verification Data** Tomasz Miksa, Ricardo Vieira, José Barateiro, Andreas Rauber
4.30–5PM	**Occam’s Razor and File Format Endangerment Factors** Heather Ryan		**The Dendro research data management platform: Applying ontologies to long-term preservation in a collaborative environment** João Rocha Da Silva, Joao Castro, Cristina Ribeiro, João Correia Lopes (Note: ends 4.50pm)
5.15PM & 5.30PM	**TOURS OF THE STATE LIBRARY OF VICTORIA** (OPTIONAL)		
7.45–10PM	**DINNER** VENUE: TO BE ANNOUNCED (NOTE: DINNER AT DELEGATES’ OWN EXPENSE)		

Thursday 9 October 2014

8AM–5PM	REGISTRATION		
9–10AM	Keynote address: The Value of Digital Preservation: Exploring the benefits of preserved data to researchers, institutions and nations Dr. Ross Wilkinson Australian National Data Service Chair: Christopher A. Lee, University of North Carolina at Chapel Hill		
10–10.05AM	HOUSEKEEPING		
10.05–10.30AM	Quick-fire Posters and Demonstrations Promo Chair: David Pearson, National Library of Australia		
10.30–10.55AM	MORNING TEA VENUE: THE COURTYARD		
11–11.30AM	Session Chair: Helen Tibbo, University of North Carolina at Chapel Hill A Perspective on Archiving the Scholarly Web Andrew Treloar, Herbert Van de Sompel	Session Chair: Andreas Rauber, Vienna Institute of Technology Building Information Modeling – A Game Changer for Interoperability and a Chance for Digital Preservation of Architectural Data? Michelle Lindlar	Session Chair: Seamus Ross, University of Toronto Supporting Analysis and Audit of Collaborative OAIS's by use of an Outer OAIS – Inner OAIS (OO-IO) Model Eld Zierau, Nancy McGovern
11.30AM–12PM	Identifying Digital Preservation Requirements: Digital Preservation Strategy and Collection Profiling at the British Library Michael Day, Ann MacDonald, Akiko Kimura, Maureen Pennock	11.30–11.50am DRM and digital preservation: A use case at the German National Library Tobias Steinke, Stefan Hein	11.30–11.50am Shaping a national consortium for digital preservation Darryl Mead
12–12.30PM	Then and Now: The Evolution of Digital Preservation and Collecting Requirements Over a Decade Leigh Rosin, Kirsty Smith (Note: ends 12.20pm)	11.50am–12.10pm Preservation of ebooks: from digitized to born-digital Sophie Derrot, Jean-Philippe Moreux, Clément Oury, Stéphane Reecht	11.50am–12.10pm The process of building a national trusted digital repository: Solving the Federation Problem Sharon Webb, Aileen O'Carroll
12.30–1.30PM	LUNCH VENUE: QUEEN'S HALL		

1.30–3PM	Posters and Demonstrations	
3–3.25PM	**AFTERNOON TEA** VENUE: THE COURTYARD	
3.30–4PM	**Session Chair: Jose Borbinha, Lisbon Technical University** **A pragmatic approach to significant environment information collection to support object reuse** Fabio Corubolo, Anna Grit Eggers, Adil Hasan, Mark Hedges, Simon Waddington, Jens Ludwig	**Session Chair: Neil Grindley, JISC** **The SCAPE Policy Framework, maturity levels and the need for realistic preservation policies** Barbara Sierman
4–4.30PM	4–4.20pm **Integrating e-government systems with digital archives** Kuldar Aas, Janet Delve, Ricardo Vieira, Ross King	**Self-assessment of the Digital Repository at the State and University Library, Denmark – a Case Study** Gry V. Elstrøm, Jette G. Junge
4.30–5PM	4.20–4.40pm **Decommissioning of legacy systems: A methodology for identifying and preserving records of ongoing business value in legacy business systems** Ingrid MacDonald, Adrian Cunningham, Anna Morris, Neal Fitzgerald	**A Digital Preservation Environment Maturity Matrix for NSLA Libraries** Sarah Slade, David Pearson, Libor Coufal
7–11PM	**CONFERENCE DINNER** VENUE: QUEEN'S HALL (NOTE: DINNER IS INCLUSIVE FOR FULL REGISTERED DELEGATES) **GENEROUSLY SPONSORED BY PRESERVICA**	

Friday 10 October 2014

8–9AM	REGISTRATION		
9–10AM	**Keynote address:** **When I say NOW, it's already over** **Herbert Van de Sompel** **Los Alamos National Laboratory** Chair: Ross King, Austrian Institute of Technology		
10–10.05AM	HOUSEKEEPING		
10.05–10.30AM	MORNING TEA VENUE: THE COURTYARD		
10.30AM–12PM	**Panel/debate:** **Are we Succeeding?** Facilitator: Shaun Hendy Andreas Rauber, Barbara Sierman, Ross Wilkinson, Seamus Ross, Ed Faye, Helen Tibbo		
12–12.20PM	**Closing remarks:** **Andrew Treloar, Australian National Data Service** Chair: Sarah Slade, State Library of Victoria		
12.20–12.30PM	**Conference close**		
12.30–1.30PM	LUNCH VENUE: QUEEN'S HALL		
1.30–4.30PM	Workshop: **Applying the TIMBUS Approach to Preserving Context in Digital Libraries** Carlos Coutinho, Paul Gooding	Workshop: **Surveying ISO Standards for PDF: archive, accessibility, engineering, metadata, 3D data and PDF itself** Duff Johnson	Workshop: **Leveraging Web Archiving Tools for Research and Long-Term Access** Lori Donovan

CONTENTS

LONG AND SHORT PAPERS

PANELS

CLOSING REMARKS

WORKSHOPS AND TUTORIALS

POSTERS AND DEMONSTRATIONS

Long and Short Papers

Linked Data Registry: A New Approach To Technical Registries

Maïté Braud
Tessella
26 The Quadrant
Abingdon Science Park
Abingdon, UK
+44-1235-555511
Maite.Braud@tessella.com

James Carr
Tessella
26 The Quadrant
Abingdon Science Park
Abingdon, UK
+44-1235-555511
James.Carr@tessella.com

Kevin Leroux
Tessella
26 The Quadrant
Abingdon Science Park
Abingdon, UK
+44-1235-555511
Kevin.Leroux@tessella.com

Joseph Rogers
Tessella
26 The Quadrant
Abingdon Science Park
Abingdon, UK
+44-1235-555511
Joseph.Rogers@tessella.com

Robert Sharpe
Tessella
26 The Quadrant
Abingdon Science Park
Abingdon, UK
+44-1235-555511
Robert.Sharpe@tessella.com

ABSTRACT

Technical Registries are used in digital preservation to enable organizations to maintain definitions of the formats, format properties, software, migration pathways etc. needed to preserve content over the long term. There have been a number of initiatives to produce technical registries leading to the development of, for example, PRONOM, UDFR and the Planets Core Registry.

However, these have all been subject to some criticisms. One problem is that either the information model is fixed and difficult to evolve or flexible but hard for users to understand. However, the main problem is the governance of the information in the registry. This has often been restricted to the host organization, which may have limitations on the investment they can make. This restriction has meant that, whilst other organizations have, perhaps, been free to use the registry they have been unable to add to or edit the information within it. The hosts of the registries have generally been receptive to requests for additions and change but this has still led to issues with timing or when different organizations cannot agree (or just utilize or interpret things in different ways).

In this paper we describe a new approach, which has used linked data technology to create the Linked Data Registry (LDR). This approach means it is simple to extend the data model and to link to other sources that provide a more rounded description of an entity. In addition, every effort has been made to ensure there is a simple user interface so that users can easily find and understand the information contained in the registry.

This paper describes what is believed to be the first linked data technical registry that can be deployed widely. The key element of the new approach is the distributed maintenance model which is designed to resolve the governance problem. Any organization hosting an LDR instance is free to add and edit content and to extend the model. If an instance of LDR is exposed on the internet, then any other organization is free to retrieve this additional information and hold it in its own LDR instance, alongside locally maintained information and information retrieved from other sources. This means a peer-to-peer network is established where each registry instance in the network chooses which other registry instances to trust and thereby from whom to receive which content. This gives control to each individual organization, since they are not dependent on anyone else but can choose to take different content from appropriate authoritative sources. At the same time it allows collaboration to reduce the administrative burden associated with the maintenance of all of the information.

General Terms

Infrastructure, Communities, Strategic Environment

Keywords

Linked Data, Digital Preservation, Automation, Technical Registries

1. INTRODUCTION

1.1 Role of Technical Registries

One of the key threats to the preservation of digital material is that “Users may be unable to understand or use the data, e.g., the semantics, format, processes or algorithms involved” [1] .

This issue is addressed in the OAIS model through the development of Representation Information networks [2]. Some of this might be specific to a given Information Object (e.g., data from a one-off experiment might need to record information related to the instrument calibration and quality control that took place) or it might apply very commonly (e.g., the need to understand the specification of PDF/A). This means that Representation Information networks will consist of some information maintained locally (to hold information specific to the Information Objects held in that repository) and some information that is probably best maintained remotely from the repository (or at least it can be done more efficiently, e.g., not every organization using PDF/A needs to be an expert in the details of its specification).

The need for Representation Information networks is well established in data-holding institutions. This is because, for example, data gathering often utilizes new combinations of techniques, methods and algorithms and thus, in order to be able to understand the results, a repository needs to be able to reference information related to these and yet does not necessarily want to repeat this information with every data set.

In memory institutions traditionally the problem has been handled in different ways using different terminology but conceptually it is the same approach. For example, usually such institutions create a catalogue entry to describe (at least at a high level) each record it holds. This catalogue entry, as well as describing information specifically about the record, may reference other information (e.g., a description of the collection to which the record belongs, or links to other controlled sources such as organizations, people or events related to the record). These controlled sources are then described in turn (externally to the individual catalogue entry) and may, themselves, reference another external source. This creates a network of information that helps a user to understand the semantics of a record.

For example, imagine a genealogist looking at the history of an ancestor. From the records of a national archive, they might be able to find out that their ancestor was in the army and served in a given regiment between two dates. The national archive might maintain a separate list of information about every regiment in the national army but might not contain detailed information about each regiment, such as where that regiment was posted on a given date. However, this information might be available from a regimental museum. Hence, a given user (with sufficient knowledge and skill) can find out where their ancestor was posted on a given date through the use of a network of representation information that will involve information held with the record, information explicitly linked to the record and information implicitly linked to the record.

For memory institutions, this sort of network applies to paper records as well as digital records and they have been in existence for some time. The advent of digital technology has made catalogues of information easier to maintain, more accessible, easier to search and easier to link to each other but the fundamental information storage and retrieval process has not changed. However, the advent of digital information has led to new problems such as the ability to continue to interpret, for example, a file of a specific format that constitutes all or part of the original record.

To solve this problem various attempts have been made to add such information to the existing, relevant representation information networks. This has included the development of ‘Technical Registries’ which are designed to be repositories of key facts about things that are important to the environment needed to interpret digital records and/or the environment needed to preserve such records.

There have been a number of high profile attempts to create such a registry including PRONOM [3], UDFR [4] and the Planets Core Registry [5]. These registries have provided significant advantages and at least some of them are in regular use. PRONOM, for example, is used as the basis for the format signatures that underpin the widely-used file format identification tool, DROID [6], while the Planets Core Registry has been used as the basis for automated characterization and migration decisions within Tessella’s digital preservation systems, SDB [7] and Preservica [8].

Other initiatives such as the "Solve the File Format Problem" [10] or the Community Owned digital Preservation Tool Registry (COPTR) [11] have already demonstrated the benefit of using crowd sourcing to collate information relevant to the Digital Preservation community but these repositories do not offer machine-to-machine interfaces and are thus aimed mainly at researchers or manual curation.

1.2 Limitations of current registries

However, all of these registry initiatives have also been subject to two main criticisms.

The first is that the set of entities modelled, the properties held about such entities and their relationship to other entities has been hard to expand and/or hard to interact with. Either of these issues makes it hard to integrate this information as part of a representation information network. For example, it would be desirable to be able to link a locally-held record about a format to, say, its formal specification. In some existing registries this could be done by, say, uploading a copy to the Technical Registry but then this would not be updated if some error was found in the specification and updated on, say, the official website.

There have been two contrasting approaches to this issue of expandability and usability. The first has been to use a fixed-schema database with a user interface intricately linked to that schema. This approach (used in PRONOM and the Planets Core Registry) makes the system easy to use but hard to expand. The alternative approach (used in UDFR) has been to use a linked data approach which is easier to expand. However, linked data is a technology designed for computer-to-computer interactions meaning that it can be hard for non-technical users to interact with the information. UDFR has made some effort to create a user interface to help with this but arguably it is harder to use the software to find information than, for example, in the fixed-schema, harder-to-expand PRONOM system.

The issue has already been raised in previous papers and initiatives such as the P2-Registry [9] recognized and proved the benefit of the Linked Data approach while highlighting that exposing SPARL query interfaces directly to end users might be too complex for a lot of people to use.

The second issue is one of governance of the information. Since these registries have been used by organizations other than their hosts, there have been issues about what to do when information is incomplete, in error or possibly subject to just being an opinion. For example, some organizations have wanted to extend the range of formats that is covered by PRONOM. The UK National Archives (the hosts of PRONOM) have been as proactive as possible at supporting such requests but the need for them to go through appropriate checks and their limited resources means that it can take some time before a request leads to a registry update. In addition, there have also been cases where there have been disagreements within the community about format definitions, and cases where an information update has changed existing behavior causing systems that relied on the previous behavior to stop working as expected.

1.3 New approach

This paper will describe a new type of Technical Registry designed to solve these problems: the Linked Data Registry (LDR). Like UDFR it uses linked data technology [12], which allows flexible linking of resources to other resources thereby offering a solution to the expandability part of the first issue.

In addition the registry aims to be as easy to search, view and edit entities as a fixed-schema system. This means it also offers a solution to the usability part of the first issue. Searches of linked data systems use a search language called SPARQL that is conceptually similar to the structured query language (SQL) used by more traditional relational databases. In many linked data systems a SPARQL end point is considered sufficient to allow for searching, viewing and editing of content. However, the users of a Registry should not be assumed to be sufficiently technically savvy to write queries using SPARQL or to be able to interpret the raw results any more than users of a traditional relational database would be expected to write SQL statements or interpret the raw results this would produce. Creating a method of allowing searching, viewing and editing of linked data information in a manner that is natural to non-technical users is a non-trivial issue that has been the subject of considerable research effort [13]. In this paper we describe how we have attempted to solve this problem. It is inevitably a design compromise but one that we believe is optimized to balance expandability and ease of use.

Crucially, LDR also addresses the issue of governance. It allows a network of registries to be created that can be replicated peer-to-peer, thereby removing the need for any organization to be dependent on any other for the maintenance of information, unless it chooses to be so.

1.4 Linked Data

Linked data is becoming a more commonly used technology but some readers may be unfamiliar with it or unclear what terminologies such as resource, subject, predicate and object mean. This section provides a very brief introduction which should be sufficient to understand the rest of this paper.

A resource is the linked data term for an entity; examples include file format, software and migration pathway. A resource needs to be uniquely identified by a URI (Uniform Resource Identifier).

A resource is described by a set of statements (expressed as subject - predicate - object). Statements can be:

- A simple statement is a statement where the object is of a simple type: e.g., a String or an Integer but not another resource
- A complex statement is a statement where the object is another resource

For example:

- “Resource A” “has MIME type” “image/jpeg”
- “Resource A” “has PUID” “fmt/44”
- “Resource A” “has extension” “JPEG”
- “Resource A” “has extension” “JPG”
- “Resource A” “has version” “1.02”

are all simple statements in the form subject - predicate - object that describe and identify resource A (aka JPEG file format v1.02).

Resource A “has internal signature” Resource B (where resource A is a file format and resource B is a DROID internal signature) is an example of a complex statement. In this case the DROID internal signature object will itself be an agglomeration of statements that define and describe it.

2. INFORMATION MODELLED

In this first version of LDR the information modelled needed to be sufficient to allow efficient (and automated) preservation-related activities to take place. However, after meeting this sufficiency criterion, the data model has been minimized deliberately.

This was partly to keep the problem tractable but also partly based on the experience of developing the Planets Core Registry. In that project we found that there was a wish to expand the data model to include every attribute that might possibly be needed in the future. This was understandable since the technology used (a relational database with a fixed graphical user interface) meant that it was hard to expand the system after it was initially completed. However, this meant in practice that large tracts of the data model were left unpopulated. Perhaps worse was that it was not clear if the lack of information meant that the data model was not useful, the information was not valuable enough to be collected, the information was too hard to collect, or maybe it had not been collected yet.

Hence, in this study, it was decided to use a technology that was much easier to expand (linked data) and to start out by only modelling the information that was known to be of interest (essentially the entities that were populated in the Planets Core Registry).

These entities could be split into two classes: factual information (information that could reasonably be expected to be held in common by lots of agencies without controversy) and policy information (information about what to do when, that might be relevant to only one repository). In LDR these two classes of information are held separately, but still linked. It should be emphasized that this is not a hard and fast distinction: just a pragmatic one. Hence, it is possible for organizations to disagree about information (such as the exact definition of a format) while it is also possible for organizations to share policies. The use of a peer-to-peer network (see section 5) allows both of these cases to be covered.

2.1 Factual Information

The Linked Data Registry (LDR) models a number of key factual entities aggregated into five groups:

- File format (with associated DROID internal signature and byte sequences)
- Software
- Related software tool (including the tool’s purpose and parameters)
- Migration pathway, including its role
- Properties and property groups

The decision to create these five groups of entities was based on how these entities are used by users. For example, a user would naturally view, create or edit information about a format and then expect to add or create an internal signature for that format. Linked data concepts mean that this relationship could be considered the other way around (i.e. internal signatures are associated with formats) especially given that a single internal signature is often associated with multiple formats. However, humans tend to look up the signatures associated with formats more often than the other way round and would tend to add new signatures based off information derived from a format's specification.

This aggregation is important for the user interface needed to interact with the system (see section 3.1 below). It is less important from a technical perspective which can safely consider the resources to be linked to each other from any perspective. The impact of this aggregation on the expandability of the model is discussed in section 4 below.

Each of these five groups of entities is now discussed in turn.

2.1.1 *Format Information*

This entity group models file formats, including internal signatures and the byte sequences of internal signatures. It is based on the model established by the UK National Archives as part of their Linked Data PRONOM research project [14].

Table 1. File Format Attributes

Attribute	**Repeatable?**	**Link to other Resource**
Identifier	N	N/A
Name	N	N/A
Version	N	N/A
Description	N	N/A
Release Date	N	N/A
Withdrawn Date	N	N/A
Internet Media Type	Y	N/A
File Extension	Y	N/A
Has Internal Signature	Y	Internal Signature
Is Rendered By	Y	Software
Is Created By	Y	Software
Is Validated By	Y	Software Tool
Has Properties Extracted By	Y	Software Tool
Has Embedded Objects Extracted By	Y	Software Tool
Has Property	Y	Property
Belongs To Format Group	Y	Format Group
Has Priority Over	Y	File Format
Has Lower Priority Than	Y	File Format
Is Previous Version Of	Y	File Format
Is Subsequent Version Of	Y	File Format

Table 2. Internal Signature attributes

Attribute	**Repeatable?**	**Link to other Resource**
Identifier	N	N/A
Name	N	N/A
Description	N	N/A
Has Byte Sequence	Y	Byte Sequence

Table 3. Byte Sequence attributes

Attribute	**Repeatable?**	**Link to other Resource**
Identifier	N	N/A
Position	N	N/A
Sequence	N	N/A
Byte Order	N	N/A
Offset	N	N/A
Max Offset	N	N/A

2.1.2 *Software Information*

This entity group simply models the existence of a software package

Table 4. Software attributes

Attribute	**Repeatable?**	**Link to other Resource**
Identifier	N	N/A
Name	N	N/A
Version	N	N/A
Description	N	N/A
Release Date	N	N/A
Withdrawn Date	N	N/A
Vendor	N	N/A
License	N	N/A
Web site	N	N/A

2.1.3 *Tool Information*

This entity group models the use of a piece of software as a tool for characterization, migration, or some other purpose. It allows modules of software packages to be specified and classified.

Table 5. Tool attributes

Attribute	**Repeatable?**	**Link to other**

		Resource
Identifier	N	N/A
Name	N	N/A
Implementation Details	N	N/A
Has Purpose	Y	Tool Purpose
Has Tool Parameter	Y	Tool Parameter
Belongs To Software	N	Software Tool
Can Extract Property	Y	Property

Table 6. Tool Purpose attributes

Attribute	Repeatable?	Link to other Resource
Identifier	N	N/A
Name	N	N/A
Applies To File Format	N	File Format
Has Priority	N	N/A

Table 7. Tool Parameter attributes

Attribute	Repeatable?	Link to other Resource
Identifier	N	N/A
Name	N	N/A
Value	N	N/A

2.1.4 Migration Pathway Information

This entity group models a migration pathway and its roles and uses.

Table 8. Migration Pathway attributes

Attribute	Repeatable?	Link to other Resource
Identifier	N	N/A
Name	N	N/A
Has Source Format	N	File Format
Has Target Format	N	File Format
Uses Tool	N	Tool
Has Target Format Group	N	Format Group
Has Validation	Y	Migration Pathway Validation
Has Role	Y	Migration Pathway Role

2.1.5 Property Group Information

This entity group models a ‘Property Group’, which is a type of information object (e.g., document, video, web site, etc.), the properties that might be expected to be measured for each such property group, and the groups of formats in which this might be manifested (called ‘Format Groups’). For example, a property group called ‘Image’ might have a series of properties (e.g., height, width, colour space, etc.) and be manifested in a whole series of ways (e.g., as a part of the TIFF format group, as a part of the JPEG format group etc.).

Table 9. Property Group attributes

Attribute	Repeatable?	Link to other Resource
Identifier	N	N/A
Name	N	N/A
Has Property	Y	Property

Table 10. Property attributes

Attribute	Repeatable?	Link to other Resource
Identifier	N	N/A
Name	N	N/A

Table 11. Format Group attributes

Attribute	Repeatable?	Link to other Resource
Identifier	N	N/A
Name	N	N/A
Belongs to Property Group	N	Property Group

2.2 Policy Information

LDR can also model policy information. In this first version this is restricted to two simple policies.

2.2.1 Tool Priority

This can be used when multiple tools are present in the Registry to carry out a task (e.g. format validation) to determine which should be used in preference to the other(s). Tool Priority is described in Tool Purpose (see Table 6) but is part of the policy section of data.

2.2.2 Migration Pathway Validation

This can be used to determine which properties should be measured before and after migration, and compared in order to check that significant properties have been maintained acceptably. It allows a tolerance to be set: for cases where the value of the significant property is allowed to change during migration.

Table 12. Migration Pathway Validation attributes

Attribute	Repeatable?	Link to other Resource
Identifier	N	N/A
Source Property	N	Property

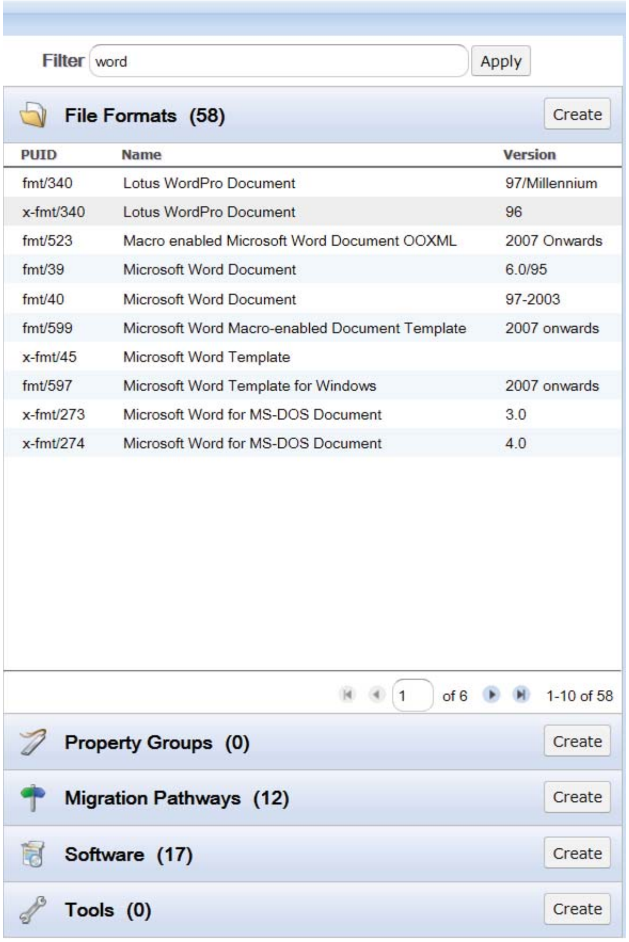
Target Property	N	Property
Tolerance	N	N/A

3. USING THE REGISTRY

3.1 Search, view and edit

As described above, one of the key features of LDR is that the registry has an easy-to-use user interface. This allows users to search for and view information about each currently supported entity. Also users with the appropriate authority can use this interface to edit information about an entity and/or add a new entity. Very importantly there is no need to understand linked data concepts or how the information is organized and stored in order to use this user interface.

This usability is achieved by using a single user interface form for each of the 5 aggregations of factual information described above (format information, software information, tool information, migration pathway information and property group information). For ease of use, the policy information is superimposed on these forms (so tool priority is displayed with the software tool entries and migration pathway priorities with the migration pathways).

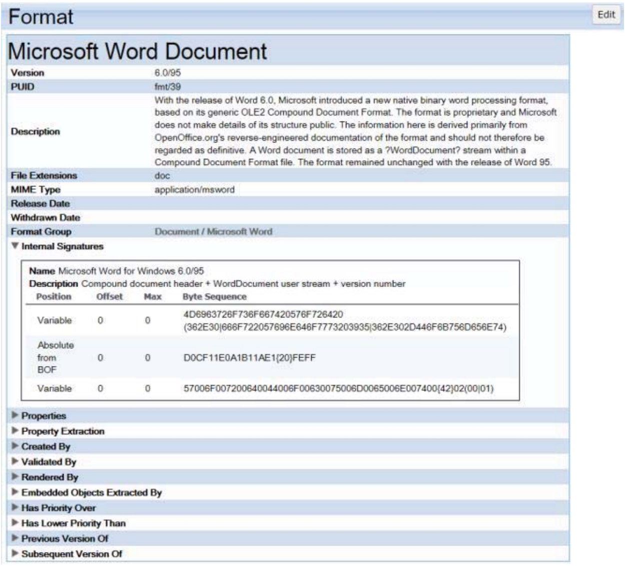


PUID	Name	Version
fmt/340	Lotus WordPro Document	97/Millennium
x-fmt/340	Lotus WordPro Document	96
fmt/523	Macro enabled Microsoft Word Document OOXML	2007 Onwards
fmt/39	Microsoft Word Document	6.0/95
fmt/40	Microsoft Word Document	97-2003
fmt/599	Microsoft Word Macro-enabled Document Template	2007 onwards
x-fmt/45	Microsoft Word Template	
fmt/597	Microsoft Word Template for Windows	2007 onwards
x-fmt/273	Microsoft Word for MS-DOS Document	3.0
x-fmt/274	Microsoft Word for MS-DOS Document	4.0

Figure 1. Simple to use search in the Registry

Rather than provide a complicated search interface, LDR allows users to filter the lists of entities in each of the 5 aggregations. There is a single filter box (see Figure 1) that filters the entity lists as each letter is typed. This makes it easy for users without training to find the information that they wish to see.

Once the user has located the information they are looking for in the relevant category, simply clicking on the item will display the information available to them. Initially the key information (e.g., name, version, identifier etc.) is shown. More detailed information (such as the internal signatures of a format including the list of byte sequences of each such internal signature) can be displayed as desired (see Figure 2).



Microsoft Word Document

Version	6.0/95
PUID	fmt/39
Description	With the release of Word 6.0, Microsoft introduced a new native binary word processing format, based on its generic OLE2 Compound Document Format. The format is proprietary and Microsoft does not make details of its structure public. The information here is derived primarily from OpenOffice.org's reverse-engineered documentation of the format and should not therefore be regarded as definitive. A Word document is stored as a ?WordDocument? stream within a Compound Document Format file. The format remained unchanged with the release of Word 95.
File Extensions	doc
MIME Type	application/msword
Release Date	
Withdrawn Date	
Format Group	Document / Microsoft Word

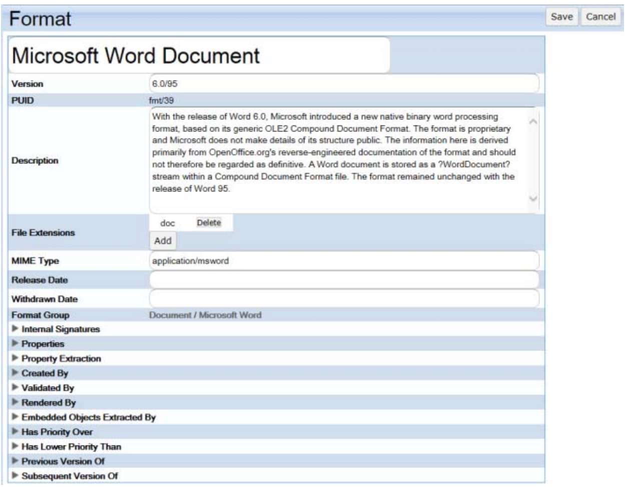
▼ Internal Signatures

Name Microsoft Word for Windows 6.0/95
Description Compound document header + WordDocument user stream + version number

Position	Offset	Max	Byte Sequence
Variable	0	0	4D6963726F736F667420576F726420 (362E30|666F722057696E646F7773203935|362E302D446F6B756D656E74)
Absolute from BOF	0	0	D0CF11E0A1B11AE1{20}FEFF
Variable	0	0	57006F007200640044006F00630075006D0065006E007400{42}02(00|01)

Figure 2. Easily understandable format information

If a user has sufficient authority to be allowed to edit information, then they can access an editable version of the user interface. This allows text to be edited, new items to be created etc. If as part of editing, a link to another resource needs to be created, then the user can choose to link to an existing resource and/or add a new resource as appropriate.



Microsoft Word Document

Version	6.0/95
PUID	fmt/39
Description	With the release of Word 6.0, Microsoft introduced a new native binary word processing format, based on its generic OLE2 Compound Document Format. The format is proprietary and Microsoft does not make details of its structure public. The information here is derived primarily from OpenOffice.org's reverse-engineered documentation of the format and should not therefore be regarded as definitive. A Word document is stored as a ?WordDocument? stream within a Compound Document Format file. The format remained unchanged with the release of Word 95.
File Extensions	doc Delete Add
MIME Type	application/msword
Release Date	
Withdrawn Date	
Format Group	Document / Microsoft Word

Figure 3. Editing format information

Each entity created by an organization will have a globally unique resource identifier (of the form: http://Creating_Organisaiton_Name/Entity_Type/Locally_Unique_Identifier).

3.2 Audit Trail

A record of every change to every resource (including its initial creation) is maintained in an audit trail. This is sufficient to allow changes to be reversed.

However, the most important aspect of the audit trail is to be able to determine which entities have been added or edited since a certain point in time. This allows different entities in the network of registries to be replicated knowing what has changed since the last such replication. The replication process is discussed in more detail in section 5.

3.3 Automation

LDR can also support key digital preservation automation features.

The first of these is the creation (and export) of a DROID signature file. This is important since it allows any organization not only to add its own formats, and their signatures, but also to be able to use DROID to identify them, even if the UK National Archives (who control the globally controlled DROID signature file) choose not to add them to their registry.

In addition, LDR also comes with the machine-to-machine interfaces needed to allow a digital preservation system to query it automatically and thereby drive decisions relating to characterization, preservation planning and preservation actions such as migration. The adequacy of this interface has been demonstrated by using it to automate preservation-related activities within Tessella's digital preservation systems, SDB [7] and Preservica [8]. This demonstrates that it is an adequate replacement for the less flexible, existing Registry previously used for this purpose (the Planets Core Registry).

4. EXPANSION

LDR has deliberately limited the initial set of modelled entities to those commonly used in digital preservation systems. Some of the existing registries support a wider data model but, as discussed above, these entities have not been heavily populated with data (if at all).

Since the new registry utilizes linked data technologies, it is easy to add resources to LDR and/or link to an external source to expand this model, if necessary. This expansion could be an additional property of an existing entity or it could be the addition of a completely new type of complex entity.

When the data model is expanded the user interface can also be expanded but, since the user interface is not created dynamically from the data model, this will take more effort. It would be possible to design a generic user interface but this would not meet one of the aims of this system: to ensure that users can easily see information and, where appropriate, add new information and edit existing information in ways that can be readily understood. We felt that a generic interface would be a big barrier to this. Hence, this is a design compromise.

LDR has been architected to offer a number of options for dealing with expansion because of this need to make a design compromise (see Figure 4). At the core of the system is a triple store with an exposed SPARQL interface. To offer a more advanced interface to client applications, there is a translation layer that combines multiple triples into more convenient to use data objects that can be accessed by such clients as either XML or RDF aggregations. The Registry user interface itself consumes these aggregations and displays the information.

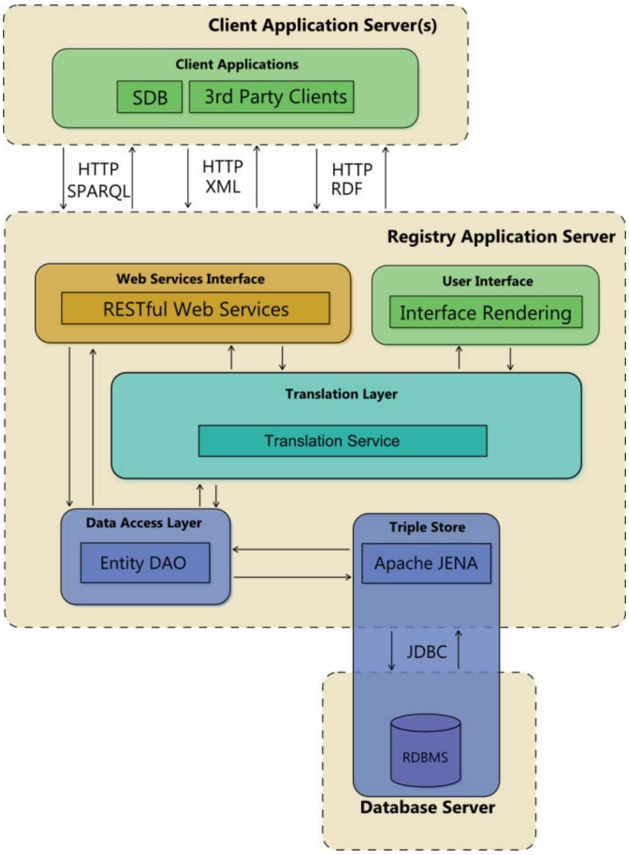


Figure 4. LDR Architecture

Hence, the options for adding new entities are (in increasing degree of effort):

- The simplest option is to just add entities to the triple store. These will be available for access by client applications via SPARQL queries and RDF.
- The next option is to, in addition to adding the entities to the triple store, enable the translation service so that the aggregations in XML can be created and validated against their XML schema. These will be available for access by client applications via a RESTful web service interface.
- The most complete option is to update the user interface as well, so that the additional information is displayed here. This could be adding additional aggregations or adding to the existing ones.

In the first version of LDR all entities in the triple store are aggregated in the translation layer and most are displayed in the user interface. It is possible that future versions will be expanded without doing this (i.e. the user interface might best be seen as a filtered view of the total information held in the triple store). It is certainly important that expansion is not prevented by the need to expand all the architectural layers.

This is an interesting design compromise that only time will tell if it has been optimized appropriately.

5. REPLICATION

5.1 Network of nodes

LDR is designed to be used as a network of registry instances, or nodes, with each node in the network being able to control its own factual information. Clearly maintaining all this information is a potentially large burden. Hence, each node can choose which node(s) to extract content (or a subset of content) from. The audit trail allows the set of potential updates since the last such extraction from a target node to be identified easily.

This means that every node can independently choose who to trust about what (and what information it wants to take on the responsibility for maintaining itself). It also means that different nodes can choose to maintain (and publish) different subsets of the total information space. These subsets can overlap with other nodes since it is up to each other node in the network to choose which other nodes to trust for which content. It does not necessarily matter if different nodes in the network hold different information about nominally the same entity, provided that the information used is appropriate to that community.

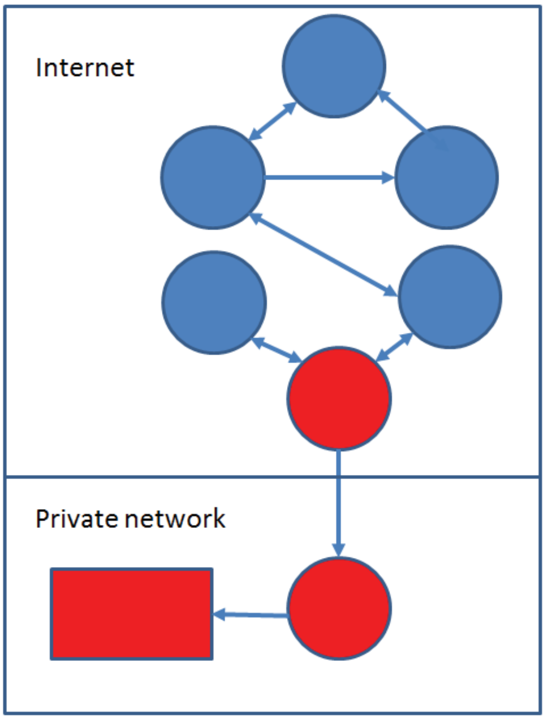


Figure 5. Possible network of nodes. Each Registry node is a circle with a rectangle representing a repository. A single organization controls the elements in red while blue entities are from different organizations.

Hence, LDR uses a peer-to-peer replication model. The advantage of this over alternative network configurations (such as ball-and-spoke, where one central node controls the content) is that it removes the need for centralized governance. Each node can control its own information and, if it chooses to, update that information immediately. At the same time the ability to extract content from other nodes means that the burden of maintaining information can be shared.

Figure 5 shows how this network could be used. In the top part of the diagram are a series of Registry nodes (each represented as a circle) in the internet which have chosen to trust all or part of the information maintained in other nodes. One organization (shown in red) is a part of this network but operates its production repository (the rectangle) inside a private network protected by a firewall. A separate (private) Registry instance serves the repository and is updated only from that organization’s public Registry instance in a controlled manner. To enable this scenario, LDR supports a data dump to enable replication without the need for a network link between nodes.

5.2 Shared instances

It is also possible for multiple organizations to share a registry instance. This allows for instance-level factual information, which would normally be controlled by a host organization (through a combination of local maintenance and choosing which other instances to trust). However, each organization using the instance could set their own independent policy information whilst sharing factual information.

6. FUTURE WORK

LDR is being rolled out first to Tessella’s customer base but then will be offered more widely. If there is sufficient interest a community version could be created.

It does bring a number of interesting challenges. It removes the need for central governance but this does not mean that there should not be guidelines for updating and adding new entities. There are likely to remain islands of excellence on which lots of other organizations will choose to depend (e.g., organizations might rely on the UK National Archives for information on standard formats as many do already via PRONOM; customers of commercial repository supplies might rely on the provider of this software for much of the information of the available tools and migration pathways used in their software etc.). It will be interesting to see who organizations choose to trust for which subsets of information and on what basis. It will also be interesting to see how organizations choose to take on the burden of the maintenance of some subsets of the necessary information themselves.

In addition, it will be interesting to see how the data model is expanded over time. We would anticipate an increase in the use of links to expand the model by linking to existing, external linked data models as opposed to adding complex new entities to the system.

7. CONCLUSIONS

Technical Registries (used to help with the preservation of digital documents, images and related content) are part of a continuum of representation information networks that include other forms of digital content and non-digital content. Some parts of this network have existed for centuries whilst others (including those covered by technical registries) are new and currently incomplete. The key lessons of existing technical registries are that:

- They must be expandable and must be able to be linked to other parts of this network.
- They must be easy to use without detailed technical knowledge.
- There must be local control of governance.

This paper describes what is believed to be the first linked data technical registry that can be deployed widely, thereby allowing the creation of a network of information maintained by a diverse and (loosely) collaborating community.

This registry has balanced the need to expand the data model with the need to make the entities in that data model easily findable, viewable and editable by non-technical users.

It establishes a replication and governance model for this network based on a peer-to-peer approach. This allows each organization to choose who to trust and which information to maintain itself. Time will tell how this new ability is utilized.

8. ACKNOWLEDGMENTS

The research leading to these results was performed as part of the ENSURE project and has received funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement n° 270000.

9. REFERENCES

[1] http://www.parse-insight.eu/downloads/PARSE-Insight_D3-4_SurveyReport_final_hq.pdf.

[2] http://public.ccsds.org/publications/archive/650x0m2.pdf.

[3] http://www.nationalarchives.gov.uk/PRONOM/Default.aspx

[4] http://www.udfr.org.

[5] http://www.openplanetsfoundation.org/planets-core-registry.

[6] http://digital-preservation.github.io/droid.

[7] http://www.digital-preservation.com.

[8] http://preservica.com.

[9] Tarrant, D., Hitchcock, S., and Carr, L. 2011. Where the Semantic Web and Web 2.0 Meet Format Risk Management: P2 Registry. In *The International Journal of Digital Curation*, Issue 1, Volume 6. DOI=http://www.ijdc.net/index.php/ijdc/article/viewFile/171/239

[10] http://fileformats.archiveteam.org/

[11] http://coptr.digipres.org/

[12] http://linkeddata.org/

[13] Davies, S., Hatfield, J., Donaher, C., and Zeitz, J.. User Interface Design Considerations for Linked Data Authoring Environments. DOI=http://events.linkeddata.org/ldow2010/papers/ldow2010_paper17.pdf.

[14] http://labs.nationalarchives.gov.uk/wordpress/index.php/2011/01/linked-data-and-pronom

New Perspectives on Economic Modeling for Digital Curation

Neil Grindley
Jisc
Brettenham House
5, Lancaster Place, London
+44 (0)203 006 6059
n.grindley@jisc.ac.uk

Ulla Bøgvad Kejser
The Royal Library, Denmark
Postbox 2149
1016 København K
45 91324747
ubk@kb.dk

Hervé L'Hours
UK Data Archive
University of Essex, Wivenhoe Park
Colchester CO4 3SQ
+44 (0)1206 822669
herve@essex.ac.uk

ABSTRACT

Society is increasingly dependent on the availability of digital information assets however the resources that are available for managing the assets over time (curating) are limited. As such, it is increasingly vital that organizations are able to judge the effectiveness of their investments into curation activities. For those responsible for digital curation, it is an ongoing challenge to ensure that the assets remain valuable in a sustainable manner. Digital curation and preservation practices are still evolving and they are not well aligned across different organizations and different sectors. The lack of clear definitions and standardization makes it difficult to compare the costs and benefits of multiple curation processes, which again impedes identification of good practice. This paper introduces a new perspective on modeling the economics of curation. It describes a framework of interrelated models that represent different aspects of the economic lifecycle based around curation. The framework includes a sustainability model, a cost and benefit model, a business model, and a cost model. The framework provides a common vocabulary and clarifies the roles and responsibilities of managers with a demand for curation of digital assets and suppliers of curation services and solutions. Further, the framework reflects the context in which managers operate and how this context influences their decision-making. This should enable managers to think through different scenarios around the economics of curation and to analyze the impact of different decisions to support strategic planning. The framework is intended to serve as a basis for developing tools to help managers analyze the costs and benefits associated with curation. The models are being developed and refined as part of the EU project 4C "Collaboration to Clarify the Cost of Curation", which is bringing together and bridging existing knowledge, models and tools to create a better understanding of the economics of curation.

General Terms

Strategic environment, digital preservation marketplace, theory of digital preservation.

Keywords

Economics, models, curation, preservation, strategy, decision-making, costs, benefits, risks, sustainability.

1. INTRODUCTION

It is difficult for organizations responsible for managing and curating digital assets to know whether they are managing those assets cost-effectively. Irrespective of the sort of data they are managing (e.g. business records, research data, cultural heritage collections, personal archives, etc.), all organizations investing in curating digital assets will expect these assets to realize some form of value over short, medium or longer timescales.

The language used to describe the management of assets over time to release value should reflect commonly used economic principles and it is through this lens that the 4C project (a Collaboration to Clarify the Costs of Curation) examined the management of digital assets and developed our framework. The framework looks at the costs of curation activities; what benefits these activities bring to stakeholders (and society as a whole); and how knowledge about these costs and benefits can help stakeholders develop sustainable digital curation strategies. More specifically though, recognizes that the management of digital assets, the realization of value, and the ability to sustain those assets for as long as needed (to realize some value) all rely on an organizations ability to make sound investments into digital curation. Or to put it another way, digital curation is the pivot around which strategic and economic planning turns and it requires a sustainable flow of resources to support it.

To ensure timely resourcing, organizations that undertake digital curation need to understand the economic lifecycle that they operate within, the costs that are incurred, and the benefits that their assets may realize. This understanding must encompass their own business processes as well as the incentives that drive funders and other stakeholders. Suppliers of asset management systems and services need to have detailed knowledge of what activities are involved, how much they cost and what the cost drivers are. They also need to understand how the systems and services generate value for their customers.

Stakeholders from the demand and supply side depend on the availability of sound financial information for accounting and budgeting. As well as knowing the factual costs, for example, records of the capital and labor costs required to develop and operate a specific system, they must also have contextual information. Context includes underlying assumptions about what is being priced, for example, the quality of the service as well as an indication of the benefits – and thus the value – that such investments represent. This financial information allows financial transactions to be recorded and analyzed for internal management purposes and may also provide greater evidence and transparency for meeting external legal requirements. It can also provide a basis for the evaluation of possible solutions and thus support budgeting and decision-making. This need for reliable and comparable financial information is exacerbated by the general growth in the amount and complexity of digital information assets that require management. This in turn puts curation budgets under pressure.

Models and tools have been developed to help organizations operate in the economic landscape and to assess the costs and benefits of digital curation. At first, interest was on assessing the costs of curation, but soon the importance of understanding the associated benefits, and stakeholder incentives for funding digital curation was also recognized by the community. This was not least owing to the extensive work of the Blue Ribbon Task Force on Sustainable Digital Preservation and Access [4]. The 4C's Economic Sustainability Reference Model (ESRM) springs from this work and aims to assist the development of sustainable strategies for digital curation [9]. Tools have also been developed to support the definition and measurement of benefits of curating research data [2,3]. An overview of models and bibliographies can be found at the Open Planets Foundation website[1], in a blog post on the Signal [13] and in a deliverable report by the 4C project [8]. A more detailed description and evaluation of current cost and benefit models can be found in another 4C report [10].

Today's trends are towards developing a unified theory of how to model the costs and benefits of digital curation in a way that will facilitate comparison of alternative scenarios and selection of good practices to ultimately gain efficiencies in digital curation [15]. Despite all the effort being put into investigations of the economics of digital curation, there is still a need to improve the map of the economic digital curation landscape and to provide practical tools that help stakeholders navigate and better understand how curation investments become more sustainable.

2. ECONOMIC MODELS

The Economic Sustainability Reference Model (ESRM) is intended to be used as a strategic tool to support planning and provoke discussion and is primarily aimed at executive and managerial staff with responsibility for managing organizational budgets rather than operational level staff undertaking curation activities [9]. The ESRM provides a foundation for the development of sustainability strategies for digital curation by organizing the problem space; providing a common reference point of concepts and vocabulary; and introducing a layer of abstraction that hides the complexities and idiosyncrasies of individual implementations and contexts, while at the same time embodying sufficient detail to support substantive discussions of shared issues.

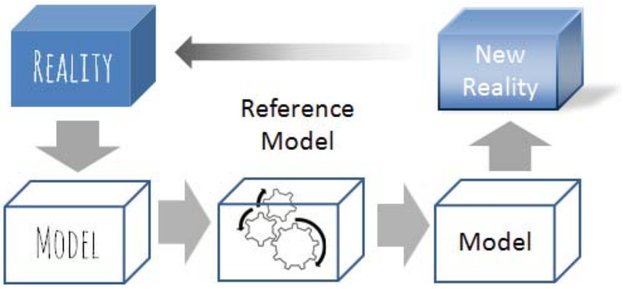


Figure 1. The purpose of a reference model.

The intention of this reference model as represented in Figure 1, is to provide people with a method of comparing current practice with an abstracted and exemplary view of alternatives; and then to provide them with an approach to advocating for change.

In relation to the modeling of digital curation, the ESRM nests within the category of economic models and is a planning resource that does not require any technical knowledge of digital curation tools and techniques. Figure 2 shows a graphical depiction of the relation between a costs model, a benefits model and an overarching economic sustainability model.

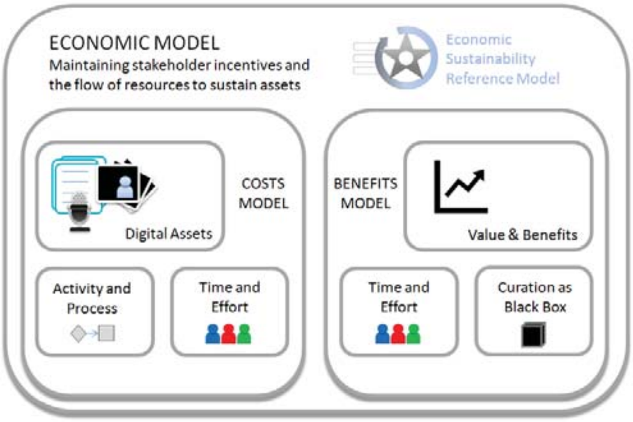


Figure 2. The nesting of costs and benefits modeling activities within the overarching framework of an economic model.

The aim of the nested model is to highlight that tackling the economics of digital curation requires a number of different perspectives and is comprised of a series of disparate tasks that occur across the curation lifecycle. Each of these tasks will be more or less achievable at different points in time depending on the organizational objectives, what resources are available to carry out curation tasks, and what information is available to help assess the potential impact of undertaking these specific tasks. At the most general economic modeling level, the motivation is to provide an understanding of why and how overall curation processes are likely to be economically affordable. This can be summarized as understanding the incentives to curate; and understanding how a flow of sufficient resources can be maintained to support these processes over time.

The ESRM maps out the key elements of the problem space planners face when designing a sustainability strategy for the digital curation processes they apply. It focuses on the general concept of a sustainability strategy, breaks it down into its key components, and draws planners' attention to the properties of those components most relevant for economic sustainability. The ESRM breaks down into four primary components:

- The Economic Lifecycle;
- Sustainability Conditions – value, incentives, selection, organization and resources;
- Key Entities - digital assets, curation processes and stakeholders (and stakeholder ecosystem);
- Uncertainties (Risks).

2.1 The Economic Lifecycle

Digital curation processes are assumed to be the central active component that require investment and are the mechanism that will ensure the sustainability of digital assets. Investment into

[1] Open Planets Foundation, Digital Preservation and Data Curation Costing and Cost Modelling, http://wiki.opf-labs.org/display/CDP/Home

curation will in turn facilitate use (or the potential for use) of digital assets and will realize value, thereby delivering a return on the investment. This could play out in a linear fashion with assets being created, curated, used and then deleted according to a retention schedule; but in the context of sustainability, it is more likely to be a cyclical process with decision points occurring from time to time when some disruption is experienced. There will be a gap in the cycle when some kind of issue (e.g. financial, technical, business, reputational) introduces an uncertainty and this will provoke a decision point, as depicted in Figure 3. The decision might be articulated as, "are we willing to change the nature of our investment to respond to the issue(s) in order to ensure the sustainability of our assets?" The decision point would more usually be prompted by a threat rather than an opportunity but it is feasible that both scenarios could be substantially disruptive in different ways.

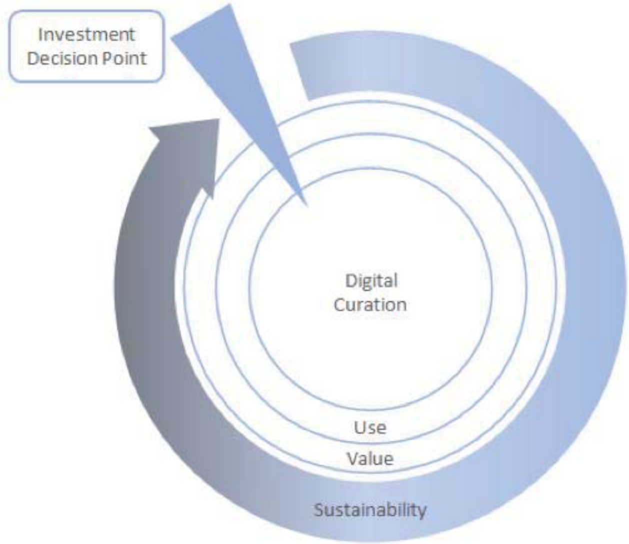


Figure 3. The ESRM Economic Lifecycle.

2.2 Sustainability Conditions

Five *Sustainability Conditions* set out issues that must be tackled to maximize the prospects for sustaining assets:

- Value – the assets must be perceived to have tangible or intangible value to relevant stakeholders;
- Incentives – relevant stakeholders must be sufficiently motivated to support and fund curation;
- Selection – where resources are scarce then discretion must be used to prioritize curation of the most valuable assets;
- Organization – the organization responsible for the curation of the assets should have an appropriate mandate; a supportive governance structure; and be optimally configured to sustain the assets;
- Resources – there must be a sufficient and ongoing flow of resources (including capital and labor) to achieve curation objectives.

2.3 Key Entities

Three *Key Entities* are proposed which are found in all digital curation contexts. Sustainability requires the nature of these entities to be understood:

- Assets – every type of digital asset exhibits various attributes or properties that to a greater or lesser extent may affect how they are curated;
- Stakeholders – the stakeholder ecosystem for digital assets can be complex and the supply side and demand side should be understood in relation to who is undertaking the curation and for the benefit of whom;
- Processes – they must be capable of (and optimized for) efficiently maintaining and possibly enhancing the value of the assets.

2.4 Economic Uncertainties (Risks)

The inclusion of *Economic Uncertainties* (Risks) is an acknowledgement that even the best sustainability strategy cannot accurately predict the future and that some expectation or mitigation of uncertainties should be built into the strategy (Figure 4).

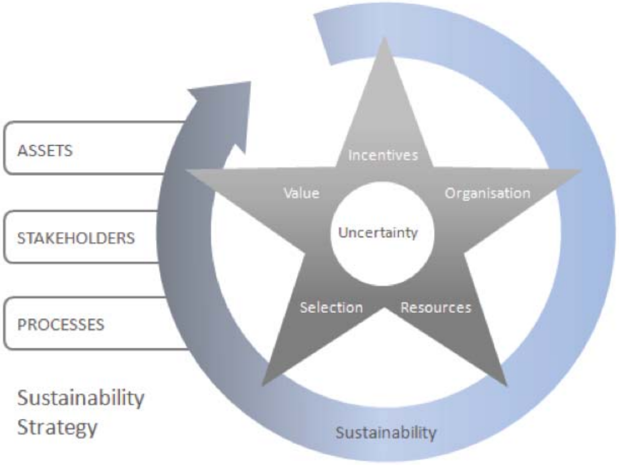


Figure 4. The ESRM components support the creation of a sustainability strategy for curation.

There is an enormous body of work on risk management and these methodologies should be employed, including the concept of negative and positive risks. Building flexibility into planning will allow the possibility of taking advantage of any opportunities that may present themselves (e.g. a cheaper service option becomes available from a different supplier; or a plan is mooted to massively upscale operations). It should also cope when a threat arises (e.g. a natural disaster substantially reduces world stocks of hard disks, or one of the major sponsors of activity unexpectedly withdraws support).

Examining the ESRM with its focus on sustainability is a useful approach to understanding the economic level of modeling, which encompasses the costs, benefits, and risks levels discussed below.

3. FRAMEWORK OF MODELS

The 4C project is developing a framework of models, terms and concepts to discuss and clarify economic decisions about digital curation and to provide common reference points. The framework is centered on the concept of the Curation Service, offered by a Provider to a Consumer (concepts are written with capital letters). The Provider and Consumer are decision-makers. Around this simple structure we then model different aspects of the economic lifecycle to explain the factors and mechanisms that impact on decision-making. The framework is shown in Figure 5.

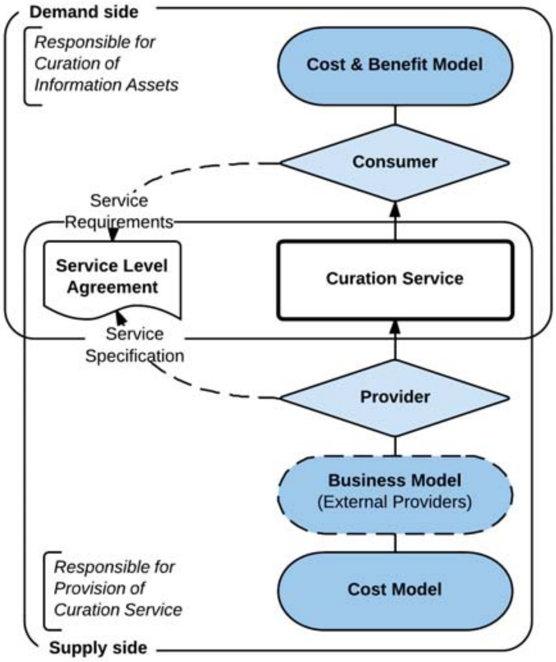


Figure 5. The 4C framework of economic models representing the demand and supply side of curation services.

The distinction between the two roles – representing the demand and supply side – is useful because the roles have different responsibilities reflecting different incentives for curation and different needs for tools. Even when services are provided in-house and the role of the Consumer and the Provider both reside within the same organization (or even with the same stakeholder group) it is useful to keep this distinction in mind when analyzing decision-making processes.

3.1 Curation Service

The Curation Service represents a value proposition; it incurs costs and should deliver benefit. It may cover the whole digital curation lifecycle or it may signify selected parts of the lifecycle, such as an ingest service or a storage solution. When it is provided in-house the Consumer can usually specify the requirements for the quality of the service – the Service Level – directly. When it comes to services that are outsourced, it may in some cases be possible for the Consumer to specify the required Service Level, while in other cases it may only be possible to select one or more predefined services.

The Curation Service can be defined in an agreement between the Provider and the Consumer, also known as the Service Level Agreement (Figure 5). Such agreements may be legally binding or have a more informal or ad hoc character, which is often the case with internal agreements, for example between two departments in an organization.

3.2 Consumer

The Consumer is responsible for the curation of information assets and must ensure that the applied Curation Service meets the organization's requirements in a sustainable way. To facilitate decision-making and strategic planning they typically use tools for costs and benefits analysis and risk management. In the framework, the demand side of the economic lifecycle is modeled by the Cost & Benefit Model.

Consumers, such as memory institutions, are of course also likely to use business models although not to address curation specifically. The value they propose to their users (and what needs to be addressed in their business case) is the services that curation enable, such as the ability to search for information assets across multiple collections. And the Cost & Benefit Model is intended to capture such benefits. Likewise, Consumers only need to know the overall costs and specifications of the quality levels of the services in order to balance cost and benefit. They see curation as a black box and do not normally need models to provide detailed cost information.

3.3 Provider

The Provider is responsible for delivering the Curation Service as agreed. The Curation Service can be supplied in-house or by outsourcing or in combination. External Providers need to generate sound business cases for services they offer, and ensure they provide return on investments (profit). Therefore, they need an exhaustive understanding of the costs associated with the services, and the cost drivers, as well as the value that the proposition brings to potential Consumers (customers). To facilitate these analyses they need business models and detailed cost models (see section 5). If the curation service is provided in-house, there may not be a need to develop a business case for curation, because the service may not be expected to realize a profit (this is indicated by the dotted line in Figure 5). In this case, Providers only need detailed cost models. The Business Model and the Cost Model represent the supply side.

Providers are also likely to use cost and benefit, and risk analysis tools, but not to optimize the curation of assets per se. Rather, these analyses are used to optimize their services, and for external Providers also their business cases and, as such, captured by the Cost Model and the Business Model.

4. COST & BENEFIT MODEL

In this section we describe the components of a conceptual Cost & Benefit Model for curation and explain how it can be used to analyze decision-making processes from the perspective of the Consumer. The model is depicted in Figure 6.

4.1 Objectives & Strategies

The Objectives & Strategies concept describes an organization's goals in terms of curation of the digital assets for which the organization, represented by the Consumer, is responsible, and outlines how it will reach these goals. The Consumer defines the Service Requirements for the Curation Service based on the Objectives & Strategies, and evaluates the Cost & Benefit of the service against these.

4.2 Organizational Context

The Objectives & Strategies are defined by the Organizational Context. Thus, Consumers make decisions in the light of the nature of the organizations and the information assets they hold, as well as stakeholders and the interests that they represent. Thus, they have to navigate a complex landscape consisting of a range of conditions where different influencers are likely to have different – and potentially conflicting – agendas. All of these intertwined internal and external conditions influence the decision-making process. To clarify the conditions we divide the Organizational Context into three key aspects:

- Organization (Mission, People, Systems)
- Information Assets (Quantity, Quality)
- Stakeholders (Internal, External)

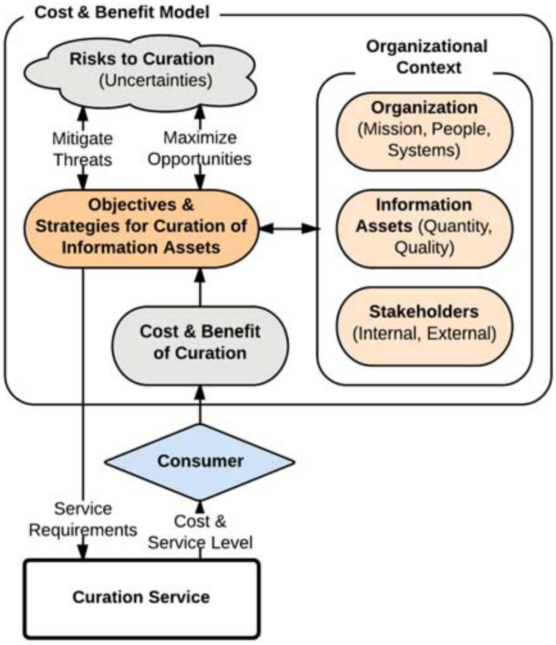

Cost & Benefit Model
Risks to Curation (Uncertainties)
Mitigate Threats
Maximize Opportunities
Objectives & Strategies for Curation of Information Assets
Organizational Context
Organization (Mission, People, Systems)
Information Assets (Quantity, Quality)
Stakeholders (Internal, External)
Cost & Benefit of Curation
Consumer
Service Requirements
Cost & Service Level
Curation Service

Figure 6. The Cost & Benefit Model for Curation represents the Consumer perspective.

4.3 Risks

The Objectives & Strategies are also influenced by Risks to Curation. This concept represents the effect of uncertainty on curation objectives. It encompasses both negative risks (threats) and positive risks (opportunities). The risks must be articulated and managed through curation strategies to minimize threats and maximize opportunities as illustrated in Figure 6. There are costs and benefits associated with mitigating or maximizing risks. The ability of a Curation Service to enhance positive risks is obviously a benefit, but so is the ability to mitigate negative risks. Again the value of the benefit will depend on the organization that the Consumer represents. If for example, an investment results in mitigation of a negative risk, this only represents value proportionally with the Consumer's incentive to reduce this risk.

4.4 Cost and Benefit

There are Cost and Benefit associated with meeting an organization's curation objectives, materialized as the Curation Service. As described above, an organization's objectives and strategies are likely to change over time influenced by its context and any risks that may be encountered. The changes further impact the requirements for services and, eventually, the Cost and Benefit of curation.

4.4.1 Costs

The costs of a Curation Service depend on which activities are included in the service and on the quality of the activities undertaken – the Service Level. Once all the involved activities have been identified and qualified, and resources attached to them, it is in principle possible to calculate the cost of the specified Curation Service. The core cost concepts needed to model these relations are described in section 5.

4.4.2 Benefits

In contrast the benefits – the advantages – of a Curation Service can only be identified and evaluated from a specific Consumer perspective. For example, if the proposed service consists of a system designed to minimize loss of data by providing multiple replicas, the perceived benefits of this service will depend on the Consumer's willingness to accept the risk of losing data. This subjective nature of benefits is illustrated in Figure 6 where the Cost & Service Level represents the information associated with the delivered service. Through the Consumer the Cost & Service Level is transformed to Cost & Benefit.

4.4.2.1 Valuation of benefits

In formal cost and benefit analysis the value of the benefits of the curation service are summed up and then the costs of providing the service are subtracted to ideally reveal the net value of the service to a given Consumer. Some benefits have a market price and it is therefore relatively easy to measure their value. Examples include the benefits of a music service that offers streaming of songs based on user fees or licenses, or the benefits of cost savings gained by investments in more efficient curation services. These benefits are also called financial or economic benefits. However, if there is no conventional market on which a benefit can be traded, no market price can be applied. It is for example difficult to assess the benefits of Europeana.eu, which aggregates European memory institutions' cultural heritage assets to make them more easily accessible to the general public, or benefits in the form of good will returned to an organization from investments in better trustworthiness of a repository. Even though, such non-financial or non-economic benefits do not have a direct market price, they still represent real value to stakeholders. Economists measure the value of benefits that do not have a market price by so-called non-market valuation techniques such as revealed preferences, which analyze past behaviors, and stated preferences (also known as contingent valuation), which asks hypothetical questions, for example about willingness to pay for a predefined change in the quality a service.

4.4.2.2 Identification of Benefits

To justify costs it is important for organizations (Consumers) to elicit and describe what the benefits of curation are, who they will benefit, how valuable they are to stakeholders, and possibly also indicate how likely it is that the benefits will realize value, and when this value will be realized. The Cost & Benefit Model provides a structure that can be used as a starting point for the identification of benefits. Thus, extending the concepts to actual instances and describing an organization's Objectives & Strategies, Stakeholders, Risks, and so on, should make it more clear to the Consumer what the benefits are.

5. BUSINESS MODEL AND COST MODEL

In this section we describe the Conceptual Cost Model (CCM) for curation and show how it relates to the Business Model. The models are depicted in Figure 7. The Business Model is not described in detail in this paper since it is still in its development phase and has not yet been fully conceptualized. Further information about the conceptual cost modeling can be found in a deliverable report by the 4C project [14].

The intention of the CCM is to provide a common foundation on which tools for assessment of curation costs can be built and to enable the specific costs of curation services and solutions to become more comparable. A concept is an abstract idea generalized from specific instances, and building on a common foundation, should enable the tools to provide comparable cost calculations at some level. The closer a tool gets to representing specific curation scenarios the more accurate the calculations are likely to be. However, the closer to specific scenarios, the less comparable the resulting cost calculations will be.

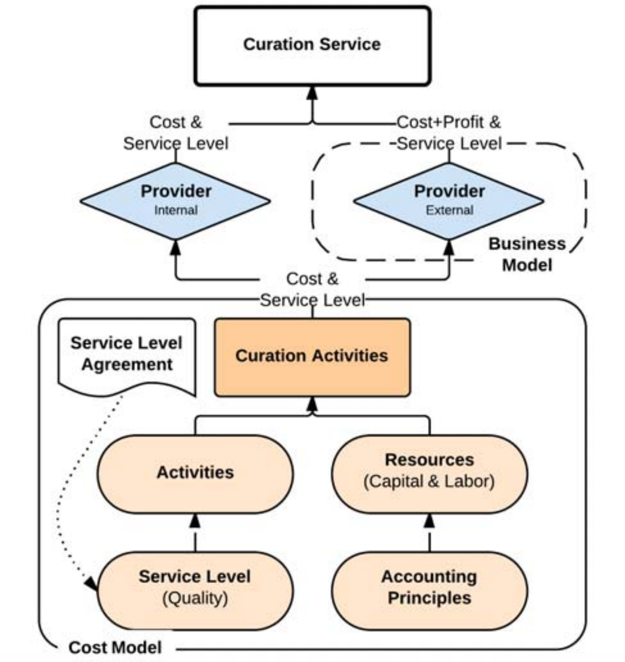


Figure 7. The Business Model and Conceptual Cost Model (CCM) represents the Provider perspective.

A cost model for curation in this context is defined as a representation that describes how Resources – direct capital and labor costs, as well as indirect costs (overheads) – required for accomplishing digital curation activities relate to costs. Cost models can further be characterized by their cost structure – the way they define and breakdown Activities and Resources, and by the way they define and handle the variables that influence the costs.

It is important for any organization providing a Curation Service to understand the distribution of costs, and what the most important curation costs are because these costs need special attention and careful management. Service Providers have to understand the factors that drive the costs up or down, such as the quantity and quality of the information assets and the length of time that the information assets will need to be curated – short or longer-term. Thus, there are many dependencies that the Provider must be aware of, for example, the costs of any systems and staff skills that are critical for delivering the service. They also need to consider how costs are likely to develop in the future, including considerations of possible financial adjustments caused by inflation or deflation. Costing digital curation is not a trivial task for a number of reasons, not least because we do not have a common understanding of the component Curation Activities [12].

5.1 Curation Activities

The costs of a Curation Service depend on the Curation Activities required to accomplish the service and on the Service Level (quality) of the activities. If the service is supplied by an external business Provider profit is normally added to the cost of delivering the Curation Service (Figure 7). Thus, the output of the CCM is a specification of the Service Level and the corresponding Cost, while the output of the Business Model, among other things, is a specification of the Service Level and the Cost including any profit.

There are many interrelated activities involved in curation and these can be implemented in many different ways and they can be set up to meet different quality requirements. This complexity makes it hard to specify the Curation Activities in a precise and clear-cut way, and it makes it difficult to delimit the costs from other business costs. Thus, there are no standardized ways of breaking down and accounting for the cost of Curation Activities. On top of this, the activities depend on constantly evolving technologies, which in turn leads to repeated changes in systems and procedures, and thus also in the costs.

5.1.1 Activities

There are numerous ways to define and breakdown activities. From the curation cost perspective we simply define an activity as a measurable amount of work performed by systems and/or people to produce a result. In order to achieve a measurement of an activity we need to break it down to a level at which we can specify the required resources, and thus get an estimate of the costs of performing the activity. The required level of granularity is also related to the required level of accuracy of the estimate.

The 4C project has used the OAIS standard [5] for a trustworthy repository as the basis for defining curation activities. The standard includes a functional model that describes a conceptual repository and three roles that interact with the repository, namely Manager, Producer and Consumer. The functional entity model divides digital preservation activities into seven functional entities: Ingest, Data Management, Archival Storage, Access, Preservation Planning, Administration, and Common Services, and these entities are further broken down in individually described functions. The PAIMAS standard [6] is an adjunct to OAIS, which provides more detailed specification of the activities around the transfer of information assets from the Producer to the repository.

Given our aim to design a generic framework to support the full breadth of possible future research and development in cost and benefit methods, we have concluded that the OAIS model, which is a well-established international standard in the field of digital preservation, provides the best starting point for breaking down Curation Activities. In fact the OAIS functional model has also been applied as a basis for the description of activities in most of the current cost models [11]. However, there are a series of challenges with applying the OAIS functional model directly to curation cost modeling.

First of all, the OAIS functional descriptions are intentionally described at an abstract and implementation neutral level. It is intended as a 'reference model'. However, costs can only be assessed against actual processes and systems. Both off the shelf services and solutions developed for specific purposes may cover multiple OAIS entities/functions or only parts of them. In these cases some mapping between the Curation Activities and OAIS entities/functions is required. Such mapping is difficult and it is further complicated by the fact that, due to the complexity of the involved activities, some of the OAIS terms are not easily understood or self-explanatory.

Second, the OAIS standard only addresses long-term digital preservation within the 'archival phase', whereas 4C aims to take a broader approach to curation such as expressed by the DCC Lifecycle viewpoint[2], which incorporates conceptualization, data creation/capture and the use and reuse of information assets.

[2] DCC Curation Lifecycle Model, http://www.dcc.ac.uk/digital-curation/digital-curation-faqs/dcc-curation-lifecycle-model

Further, it also applies to organizations and projects with a remit limited to short and medium term storage. Thus, curation covers the full lifecycle of information assets, and these activities may be expressed by the three OAIS roles covering production, use, and management activities in addition to the repository activities.

In conclusion we have decided to use the OAIS standard to populate the activity model in the framework as far as possible, but we also acknowledge there may be a need to bend the standard in some ways to make it more applicable to costing. Any such amendments would need to be justified by a particular curation cost model developer. The proposed framework extends to support the full curation lifecycle and divides activities in levels, starting from the high-level roles, functional entities and functions, which are used by the OAIS standard and, if required, allowing for further breakdown of OAIS functions into measurable activities.

The activity breakdown structure includes the following entities:

- Production: including for example conceptualization, creation of information assets, capture, and digitization
- Pre-ingest: including for example appraisal, selection, and preparation for ingest
- Ingest: ingest of information assets
- Storage: short and long-term storage and maintenance of information assets
- Data Management: management of descriptive and administrative data
- Access: provision of access to information assets
- Lifecycle Planning: planning, research and development of curation activities
- Administration: administration of repository systems, standards and policies
- Common services: including services necessary to support a repository such as inter-process communication, name services, temporary storage allocation, exception handling, security, and directory services
- Use: use and re-use of information assets, including for example interfaces for crowdsourcing
- Management: including for example the provision of overall budgets and policies, and any certification related activities

5.1.1.1 Service Level

The Service Level defines the quality of the Activities. It is usually specified in a Service Level Agreement (Figure 5 and 7). The lack of a clear way of defining and measuring Service Levels represents an important challenge in cost and benefit modeling because of the close relationship between the Service Level of the Curation Activities and the Cost, as well as between the Service Level of the activities and the Benefits perceived by the Consumer. If for example we consider the activity to 'store information assets' the Service Level of the activity may among other things specify that three copies of the assets are stored. All other things being equal, the Cost of this activity will be proportional to the number of copies specified. Likewise, the number of copies will normally be proportional with the level of information integrity because the more copies the lower risk of data loss. However, it will be inversely proportional to the level of confidentiality because the more copies that exist, the higher the risk of compromising access. Therefore, the same Service Level (quality) of the activity may have different value to different Consumers, depending on the Service Requirements in relation to costs, integrity, and confidentiality.

The Service Level may be evaluated through quantitative (e.g. pass/fail, minimum score, certification level) or qualitative measures (such as descriptions of the quality). Thus, the Service Level can be a defined quality criteria for an activity; a more complex and formal agreement between two or more units; or a higher level of service 'quality' formalized through a certification process, for example through ISO 9000[3] or ISO 27000[4]. There are also more or less standardized ways to certify the quality of repositories for long-term preservation and access. For example, ISO 16363 [7], Data Seal of Approval (DSA)[5], Trustworthy Repositories Audit & Certification: Criteria and Checklist (TRAC)[6], Information and Documentation - Criteria for Trustworthy Digital Archives (DIN 31644)[7]. These audit and certification instruments can help to establish quality measurements.

5.1.2 Resources

Activities are performed both by systems and people. Thus, to complete an activity a certain amount of resources are required, and for accounting purposes these are often divided into Capital and Labor costs. Resources are what must be expended to deliver activities.

Capital Costs include, for example, building space (server space, office space, and so on), equipment (servers, network, and the like), energy (for systems, cooling, et cetera) and materials (storage media, and so on). Depreciation (for tangible assets) and amortization (for intangible assets) are mechanisms for distributing capital costs over the estimated useful lifetime of an asset to indicate how much of an asset's value has been used. For example, the time in which a server becomes obsolete may be five years. With a 5-year time period the cost of using this resource will be its acquisition cost, whereas with a 1-year period the cost would be the depreciated acquisition cost.

Labor costs consists of salaries and any benefits paid to staff for a period of time or for a certain job. Salaries are normally differentiated by job functions (developer, metadata officer, etc.) and possibly also by skill level, seniority and/or performance. The labor costs required to complete an activity can be expressed as a monetary value – the cost of salaries multiplied by time expended on the activity – but they may also be expressed simply in time – as the time it takes to complete the activity for a certain job function. The advantage of measuring labor costs in time is that it makes the figures more comparable across organizations and countries, where there may be significant differences in salaries. If needed the time measure can be translated into monetary values for a specific scenario. If for example the cost of running a system

[3] ISO 9000 Quality Management, http://www.iso.org/iso/home/standards/management-standards/iso_9000.htm

[4] ISO 27000 Information Security Management Systems, http://www.iso.org/iso/home/store/catalogue_ics/catalogue_detail_ics.htm?csnumber=63411

[5] Data Seal of Approval (DSA), http://datasealofapproval.org/en/

[6] Center for Research Libraries (CRL), http://www.crl.edu/archiving-preservation/digital-archives/metrics-assessing-and-certifying/trac

[7] Nestor Seal for Trustworthy Digital Archives, http://www.langzeitarchivierung.de/Subsites/nestor/EN/nestor-Siegel/siegel_node.htmltml

takes a developer 20 hours per week, this figure can be multiplied with salaries applicable to the job functions in different countries. Along this line, the unit Full Time Equivalent (FTE) is used to make workloads comparable. FTE expresses the workload as the ratio of the total number of working hours during a certain period by the number of full-time working hours in that period. 1 FTE is equivalent to that of a person working full time for a year.

Capital and labor costs can also be divided in direct and indirect costs. Direct costs are those directly used for performing digital curation activities, such as costs of acquisition of storage media or the costs of staff employed to add metadata. Indirect costs, also called residual costs or overheads, are those incurred by the usage of shared resources, such as general management and administration or common facilities and systems, where it has not been feasible to allocate the cost to specific activities.

Variable costs fluctuate depending on the amount of activities being undertaken and are differentiated from fixed costs, which do not depend on the amount. For example, the cost of materials used to complete an activity is a variable cost, as opposed to salaries and rents, which are fixed regardless of the amount of activities. Thus, variable costs are normally equal to direct costs and fixed costs to indirect costs. However, given enough scale and time, no cost is really fixed.

Costs can also be divided in one-time costs, periodic (term) costs or recurring costs, depending on the time period. The term capital or investment cost is often used to denote a one-time cost incurred upon the acquisition of equipment such as a storage system. The term periodic cost is used to indicate that the cost will be incurred at irregular intervals. Recurring costs, also known as running costs or operating costs, include costs relating to the consumption of media, energy and labor.

Other important time related aspects of costs include inflation (general price increases), individual price changes that are related to specific resources – such as storage media, energy, office space, computer scientist wages – and interest, which reflect economic growth and cost of capital. Even though the cost of resources has in general been increasing, the cost of both capital and labor per unit of digital information assets has, due to technological innovation, been decreasing over the past decades (although at very different rates). Therefore, in order to calculate the present value of estimated future costs different discount rates are preferable. The present value is needed in order to compare different cost scenarios over time.

Costs can be divided by accounting periods to capture past cost (ex post) and/or future costs (ex ante). Records of past cost are used in accounting whereas estimations of future costs over certain time periods (such as months, quarters, and years) are used for budgeting.

5.1.2.1 Accounting Principles

Accounting can be defined as a set of concepts and techniques that are used to measure and report financial information about an economic unit [16]. In order to make financial reports understandable and comparable between organizations, the reports need to follow generally accepted accounting principles (GAAP) defined by national and international standardization bodies. The International Financial Reporting Standard (IFRS) Foundation is an independent, not-for-profit private sector organization working in the public interest to develop and promote the use of a single set of globally accepted, international financial reporting standards through its standard-setting body the International Accounting Standards Board (IASB)[8]. Thus, the Accounting principles, delivered as national or international standards should govern standard accounting practices.

Just as it can be difficult to segregate the costs, which are incurred when carrying out Curation Activities it can be difficult to segregate costs that are incurred within Resources. The Transparent Approach to Costing (TRAC)[9], which is applied in Higher Education in the UK, has been suggested as a concrete tool for recording resource cost data in relation to research data [1].

6. DISCUSSION

The approach taken has been to accept that the models have different purposes (communication, simplification, common understanding of basic relationships, complex expression of curation concepts in a specific context) and that where there are overlaps, either in purpose or terminology, perfect interaction and synchronization between them will not always be apparent. But the important factor is to understand that no particular approach or view of a system exists in isolation and that, where possible, models should be designed and expressed within the context of the higher level and more granular surrounding models. The ESRM and the framework help to clarify and signpost these relationships.

The establishment of the framework with its distinction between those with a demand for curation of assets and those that supply curation services has enabled us to clarify roles and responsibilities at the conceptual level, namely that of the Consumer and the Provider. The distinction may seem rigid and indeed in real life roles are often less clearly defined, but it has proved useful for identifying the kind of models and tools that are required to support decision-making related to the economics of digital curation. Further, it has been useful for clarifying the relationships between the different models (Cost & Benefit, Business Model, Cost model), as well as to define the kind of financial information the models deliver.

On the demand side we found that to ensure that the information assets remain sustainable Consumers basically need tools for analyzing the cost and benefits of Curation Services. This includes the ability to assess the cost and benefit of alternative services and of managing risks. As a first step to facilitate such analyses the Cost and Benefit Model defines and describes – at a conceptual level – the dynamics of the determinants that influence the costs and benefits of curation including risks. The model is still under development, but we have shown how it may already help identify potential benefits of curation.

On the supply side we found that Providers need tools that will help them assess how the costs vary with the quality of the service being applied. To this end it became clear that it is also necessary to distinguish between internal and external Providers. The reason is that the latter need business models in addition to cost models, to generate profitable business cases.

An ongoing challenge is the tension between the need for very specific local application of terms and concepts and the need to have common terms and classifications if models and their outputs are to be more generally understood and ideally comparable. These tensions between generally applicable and

[8] International Financial Reporting Standards (IFRS) Foundation, http://www.ifrs.org/The-organisation/Pages/IFRS-Foundation-and-the-IASB.aspx

[9] TRAC, http://www.jcpsg.ac.uk/guidance

understood concepts and the need for local specifications apply throughout complex systems of all types. There is not yet any authority to yield a ‘big stick’ when encouraging the use of standardized terms and classifications, And only by researching, defining and presenting likely ‘controlled vocabularies’ and promotion of the benefits of their re-use will we see the slow agreement and use of common definitions.

The framework we have described here is conceptual. There are more advantages of describing the models at a conceptual level. First of all, it provides a common framework for defining the cost and benefit of a curation service unambiguously which is a prerequisite for making cost and benefit comparable across different scenarios. At the conceptual level the model should in principle be able to encompass all use cases and in this sense it may serve as a guide for developers of cost and benefit models.

Also the concept models supports the clarification of central economic terms and encourages a common language around costs and benefits, and in this way it also supports communication and exchange of knowledge. The lack of a universally accepted terminology and clarification of cost and benefit concepts has previously been shown to be an important obstacle for reaching consensus on how to model these [11]. The 4C project is developing a Curation Costs Exchange platform (CCEx) where cost data and information about the cost data can be shared[10]. A key aim for CCEx is to employ standard use of terms and classifications.

Given the complexity of assessing costs and benefits and the entailed complexity of any tool aiming to simulate this complexity, it is unlikely that any single tool will be able to handle all scenarios. However, it may be realistic that tool developers can use the concept model as a basis to ensure that the resulting assessments are comparable, and then develop tools on top of the model for different groups of similar stakeholders (profiles). It should be possible for developers of cost and benefit tools to interpret and populate the concepts according to the context they need to address whilst maintaining references to more generic elements. This should make it possible to provide financial information that maps onto comparable entities, which in turn may mean that profiles for specific types of organizations working in similar environments can be developed.

Tackling complexity by the application of detailed models is likely to come with increased costs of collecting the required cost data and information, and these costs must be justified by a correspondingly greater utility of the results. So it is important for users to define the purpose of the modeling in order to understand their requirements in terms of the degree of granularity and accuracy that they will expect the model to deliver. The process to define activities is in general beneficial to any organization since it will improve their understanding of the activities and workflows and allow for possible optimizations.

We have decided to base the generic CCM on the functional model defined in the OAIS standard. Even though OAIS is a reference standard and does not define the entire digital curation lifecycle it is still the most detailed and widely used standard that relates to the field of digital curation. However, in order to encompass curation scenarios other that those for long-term trustworthy preservation, there is a need to relax some of the requirements, for example, to encompass scenarios where information assets only need to be retained for the short or medium term.

Extensions of the OAIS model to cover the full lifecycle are critical to the remit of 4C and curation costing in general, as are exceptions which support those with responsibility for storing information assets over the short and medium term (e.g. encompassing storage as well as full archival storage) but until the OAIS has been specifically researched and found appropriate for cost-assignment, or a commonly accepted alternate approach has been developed, these core functional entities should remain our common benchmark and deviations from that benchmark should be documented and justified when applied to a particular curation cost methodology. These may be primarily for practical reasons such as dividing the more esoteric costs of planning, management and administration into more direct cost centers such as production, ingest, storage and access.

Similarly maintaining a clear link between terminology and the OAIS benchmark and those used in a particular approach will support the ongoing comparison of approaches. This will help to drive adoption of a common approach by defining how the model and specification should be updated over time to take account of changes in the broader environment.

7. CONCLUSIONS

In this paper we investigate the usefulness of new approaches to modeling the economic landscape of curation and have set out a nested Economic Sustainability Reference Model, which indicates some hierarchy of scope. An economic level of modeling is the broadest and most encapsulating activity and subsumes not only all of the other approaches referenced in this paper but also has a relationship with business models. This has not been touched upon in detail here but is, in fact, being addressed by ongoing work on the 4C project. Sustainability planning is proposed as a form of economic modeling and one that can largely stand in to represent how to think about digital curation from an economic perspective. ‘Largely’ rather than ‘wholly’ to acknowledge the gap left by business planning and the related analyses and assertions that would form part of that process.

The next nested layer focuses on costs and benefits modeling considered as a dual concept and providing a framework for sensibly informing decisions that may need to be taken in relation to adopting or rejecting curation services.

Also we have described a framework of conceptual models, including a Cost & Benefit Model, a Business Model and a Cost Model focusing on the roles and responsibilities of the Consumer and Provider of Curation Services and shown how it can help clarify decision-making processes. More specifically it has clarified the relation between the models and their outputs. In addition, it has highlighted that while the costs of curation can in principle be assessed objectively once you have identified the activities involved and the resources required to complete them, the value of benefits of curation can only be assessed in relation to a specific stakeholder.

The work set out in this paper leads to some conclusions about future work and much of this follows from the points made above (see section 6 - Discussion).

* This is a complex area and there is further work to do to adequately join up existing models and to define new ones that will help to make sense and provide a more coherent perspective on the economics of digital curation;

[10] Curation Costs Exchange (CCEx): http://www.curationexchange.org

* Related to that complexity, a lot more work needs to be done to standardize terminology and all types of modeling (economic, costs, benefits and business) need further validation from diverse groups of stakeholders;

* The OAIS is an imperfect foundation for breaking down activity-based costing approaches but it is the only real practical and widely accepted standard that can currently be referenced.

Looking specifically at two of the diagrammatic representations in this paper (Figure 3 and Figure 6) another conclusion that presents itself is the importance of the decision-making moment as a fundamental design feature of economic modeling.

It is also clear, in terms of the work that the 4C project has done, that the models and other resources are beginning to usefully join up concepts and link the whole area together but there is a great deal more work that can now more clearly be set out. This can usefully be described and addressed by the 4C Roadmap [17], which will be the final output of the project and will synthesize all of the learning and conclusions into an action agenda for the wider community.

8. ACKNOWLEDGMENTS

We thank 4C Project colleagues who have done extensive and collaborative work on many aspects of this paper.

The 4C project is funded by the European Commission as part of the FP7 ICT Cooperation programme, contract 600471.

9. REFERENCES

[1] Beagrie, N., Chruszcz, J., and Lavoie, B. 2008. *Keeping Research Data Safe: A Cost Model and Guidance for UK Universities*. JISC, p.13, http://www.jisc.ac.uk/publications/reports/2008/keepingresearchdatasafe.aspx

[2] Beagrie, N. 2011. *User guide for keeping research data safe. Assessing costs/benefits of research data management, preservation and re-use*, Version 2.0., Charles Beagrie Limited, p.24-31, http://www.beagrie.com/static/resource/KeepingResearchDataSafe_UserGuide_v2.pdf

[3] Beagrie, N. and Houghton, J. W. 2014. *The Value and Impact of Data Sharing and Curation: A synthesis of three recent studies of UK research data centres*, Jisc., p. 8-11 http://repository.jisc.ac.uk/5568/1/iDF308_-_Digital_Infrastructure_Directions_Report%2C_Jan14_v1-04.pdf

[4] Blue Ribbon Task Force on Sustainable Digital Preservation and Access. 2010. *Sustainable Economics for a Digital Planet: Ensuring Long-Term Access to Digital Information*, Final report, http://brtf.sdsc.edu/biblio/BRTF_Final_Report.pdf

[5] CCSDS (Consultative Committee for Space Data Systems). 2012. *Reference Model for an Open Archival Information System (OAIS)*, CCSDS 650.0-M-2, Magenta Book, (ISO14721:2012), http://public.ccsds.org/publications/archive/650x0m2.pdf

[6] CCSDS (Consultative Committee for Space Data Systems). 2004. *Producer-Archive Interface Methodology Abstract Standard (PAIMAS)*, CCSDS Magenta Book, (ISO 20652:2006, reviewed and confirmed in 2009) http://public.ccsds.org/publications/archive/651x0m1.pdf

[7] CCSDS (Consultative Committee for Space Data Systems). 2011. *Audit and Certification of Trustworthy Digital Repositories*, CCSDS 652.0-M-1, Magenta Book, (ISO 16363:2012), http://public.ccsds.org/publications/archive/652x0m1.pdf

[8] Ferreira, M. and Farier, L. 2013. *Baseline Study of Stakeholder & Stakeholder Initiatives*, Deliverable report D2.1, Collaboration to Clarify the Costs of Curation (4C), http://www.4cproject.eu/d2-1-stakeholders

[9] Grindley, N. and Lavoie, B. 2013. *Draft Economic Sustainability Reference Model*, Milestone report, Collaboration to Clarify the Costs of Curation (4C), http://www.4cproject.eu/ms9-draft-esrm

[10] Kejser, U.B. et al. 2014. *Evaluation of Cost Models and Needs & Gaps Analysis*, Deliverable report D3.1, Collaboration to Clarify the Costs of Curation (4C), http://www.4cproject.eu/d3-1

[11] Kejser, U.B., et al. 2014. State of the Art of Costs and Benefit Models for Digital Curation. In *Proceedings from IS&T Archiving Conference*, Archiving 2014 Final Program and Proceedings, ISSN 2161-8798, Online ISSN: 2168-3204, 144-149, http://ist.publisher.ingentaconnect.com/content/ist/ac

[12] Kejser, U.B, Nielsen, A. B., Thirifays, A., 2012. Modelling the Costs of Preserving Digital Assets, In *Proceedings of UNESCO Memory of the World Conference "The Memory of the World in the Digital Age: Digitization and Preservation"*, Eds. L. Duranti, E. Schaffer, 529-539, http //www.unesco.org/new/fileadmin/MULTIMEDIA/HQ/CI/CI/pdf/mow/VC_Kejser_et_al_27_B_1350.pdf

[13] Lazorchak, B. 2014. A Digital Asset Sustainability and Preservation Cost Bibliography, *The Signal – Digital Preservation*. Blog post. http://blogs.loc.gov/digitalpreservation/2014/01/a-national-agenda-bibliography-for-digital-asset-sustainability-and-preservation-cost-modeling/

[14] L'Hours, H. et al. 2014. *Cost Concept Model and Gateway Specification*, Deliverable report D3.2, Collaboration to Clarify the Costs of Curation (4C), http://www.4cproject.eu/d3-2-ccm-grs

[15] Lungi et al. 2012. Economic Alignment. In *Aligning National Approaches to Digital Preservation*, Ed. N. McGovern, 195-268, http://educopia.org/publications/ANADP

[16] Walther, L. M. 2013. Welcome to the World of Accounting. In *Financial Accounting*. http://www.principlesofaccounting.com/chapter1/chapter1.htm

[17] 4C Project, 2014. *Investing in Curation: A Shared Path to Sustainability.* 4C Draft Roadmap. http://4cproject.eu/d5-1-draft-roadmap

Achieving Canonical PDF Validation

Duff Johnson
PDF Association
Neue Kantstrasse 14
14057 Berlin, Germany
+1 617 283 4226
duff.johnson@pdfa.org

ABSTRACT

While PDF is the best currently available option for archiving fixed-form electronic documents, low quality PDF files remain problematic throughout the document lifecycle, and can pose substantial challenges for memory institutions.

This paper proposes a model for realizing and promulgating PDF validation based on a canonical (i.e, accepted industry-wide as definitive) approach rather than focusing on preservation per se.

General Terms

strategic environment, preservation strategies and workflows, specialist content types, digital preservation marketplace

Keywords

PDF, PDF/A, software, validation, standard, canonical, adoption

1. INTRODUCTION

The Portable Document Format (PDF) was invented by Adobe Systems and first released with Adobe's Acrobat software in 1993. The value proposition was simple: reliability when shared. PDF has largely delivered on that promise - but not entirely.

21 years later PDF is an ISO standardized format. For electronic documents, PDF is an exemplar *de facto* standard as well [2].

This paper proposes development of a canonical (accepted industry-wide as definitive) validation model encompassing all PDF features and thus enforceable across the document lifecycle.

2. PDF RISES

For printing technology vendors PDF's popularity began to take off with the November 1996 release of PDF 1.2, but marketplace uptake was slower. The early PDF specification was too flexible; reliability was hard to guarantee. Workflows suffered when users encountered formally "valid" files too difficult (or impossible) to process [12]. The problems were serious and widely felt.

The industry's successful response was PDF/X, a subset of PDF designed to ensure reliable exchange in prepress workflows. PDF/X became the first ISO standard for PDF technology [17].

As PDF became popular for printing formal documents, the use of PDF for distribution and retention of electronic documents became commonplace in every functional area within business and government organizations, and as part of website content.

2.1 PDF/A (archive) and ISO standardization

PDF/X was not a general-purpose standard. Responding to industry and governmental requests for PDF files suitable for long-term retention, industry stakeholders and trade groups began development of a PDF subset for archival-grade electronic documents. In the PDF context, "archival-grade" means embedded fonts, no external dependencies and prohibition of certain functionality such as encryption and JavaScript. ISO 19005 (PDF/A-1) was published in 2005 and has been adopted by the US National Archives and Records Administration (NARA) [16], and by governments and businesses worldwide [18].

Other ISO-standardized subsets of PDF have followed: PDF/E for engineering, PDF/UA for accessibility and PDF/VT for variable data and transactional printing.

In the mid 2000s Adobe Systems realized that turning over PDF to the ISO was the right move to drive continued adoption of the format. While the PDF specification was freely downloadable and PDF viewer software is traditionally free, the fact that PDF was proprietary inhibited governments, engineering concerns and other preservation-minded institutions from comfortably standardizing their own publishing, accounting, enterprise content management (ECM) or line of business (LOB) systems on PDF.

With thousands of implementers and worldwide acceptance, PDF had become "too big to own". In the spirit of the company's original - and commercially brilliant - move to publish PDF's specification for free, Adobe offered PDF to the ISO for open, democratic management by a committee of volunteer experts. PDF 1.7 was thus standardized as ISO 32000-1 in 2008 [17].

2.2 A *de facto* standard for electronic paper

From the end user perspective PDF serves as electronic paper. Self-contained, reliable, flexible and resolution-independent, PDF is easy to make from any electronic source, and freely viewable on any platform. Emulating many key characteristics of paper has helped make PDF the most popular format worldwide for downloadable electronic documents [5] [17]. The format's nearly universal adoption makes it typical in common use-cases:

- Print or distribute finalized documents – "Post the PDF"
- Retain, share or manage draft documents - "PDF it"
- Annotation of 3rd party content – "Add a note to the PDF"
- Capture content from arbitrary source - "Scan to PDF"
- Collate from arbitrary sources – "Insert / replace PDF pages"
- End-user data capture – "Fill the PDF form"

Worldwide implementation and use of PDF technology shows no sign of abating; searches for PDF files continue to increase over time, and in contrast to other formats [7]. On the public internet, institutions communicating on formal terms tend to be heavy users of PDF [8], while privately held transactional and other documents in PDF are estimated to be in the billions [20].

2.3 Beyond the printable page

While PDF is fundamentally a page description model for text, vector graphics and bitmap images, the technology includes many features that distinguish it from image formats such as TIFF and JPEG. Increased utilization of document-oriented features such as forms, annotations, XMP metadata, digital signatures, encryption, 3D, geospatial, video, embedded files, tagging and other advanced capabilities represents a growing challenge for the preservation community – a challenge that existing tools and workflows do not address in a cost-effective manner. Meanwhile, the volume of content that meets retention criteria is exploding [14].

3. CHALLENGE AND OPPORTUNITY

Although end users have enthusiastically adopted PDF the digital preservation community is more circumspect. Although research libraries prefer PDF/A to formats such as HTML or RTF they rate PDF itself as only slightly preferable to HTML [19].

In addition to concerns over PDF's complexity and reliability, some features that help make PDF compelling to end users complicate efforts to ensure electronic content remains accessible in the future. As a result, although PDF is generally extremely reliable and accepted in the marketplace, archivists have hesitated in trusting PDF as a long-term storage format [1].

An opportunity exists to harmonize industry's interest in promoting investment in PDF technologies with archivists' interest in reliable files, low-cost ingestion and maximum longevity. In the next section this paper provides an overview of the historical and technical reasons for archivists' concerns before moving on to discuss solutions as seen through an industry lens.

4. THE PROBLEMS

Compared to HTML PDF is a very complex file format. It includes 11 syntaxes, at least 20 native and 3rd party binary formats, 10 stream filters, 2 encryption algorithms, and more. Beyond PDF's rich imaging model the format includes interactive forms, encryption, digital signatures, annotations, embedded files, accessibility features and more [11].

The challenges PDF technology presents to archivists may be organized into five categories:

1. **Complexity.** Compared to plain text or TIFF, PDF is technically complex.
2. **PDF has changed.** While remaining backwards compatible, the PDF specification has changed (it has become more detailed and rigorous, as well as richer) over time. Even so:
    a. Old and "flaky" PDF files exist.
    b. Old software is still making flaky PDF files.
    c. Good files can be damaged by old or bad software.
3. **Varying degrees of support.** Few implementers claim to support all the functionality defined in PDF, which is fine. However, many implementers do not fully address the features they do claim to support.
4. **Fonts.** ISO 32000 does not require embedded fonts, so it is possible to inadvertently create unreliable PDF files.
5. **No canonical model for validation.** Today, developers must rely on their own tools or open-source applications lacking broad industry acceptance such as JHOVE to identify potential problems. Unfortunately, it is often difficult to determine with certainty whether or not a problem even exists. What should a digital preservation professional do if a PDF fails JHOVE, but passes Adobe's Preflight? Adobe Reader is not useful as a validator precisely because it is designed to accommodate very poor quality PDF files.

5. THE SOLUTIONS

The technical problems are significant, but the scope and scale of any given software development project may be the least of the barriers to addressing archivists' concerns about PDF.

Billions upon billions of PDF files already populate the world's desktops, shared-drives, ECM systems, SharePoint servers and websites. Obsolete software cannot be willed out of existence. Enforcement of policies, from embedded fonts to embedded files, will not occur spontaneously.

Let's review notionally and practically plausible responses in each problem category, looking for common threads.

5.1 Problem 1: PDF is technically complex

This problem is fundamentally ineradicable, since any other self-contained file format – even one designed only for rendering – would have to be similarly complex, at least in contrast to bitmap or ASCII-based formats that cannot replace PDF's functionality.

Solution: Developer education, ideally, via tools that deliver canonical information, analysis and advice about input PDF files.

5.2 Problem 2: PDF has changed over time

PDF was born as little more than a page description model, but it evolved through contact with the marketplace. Today's PDF (ISO 32000-1) has far more features compared to PDF 1.0, including support for rich content, transparency, new font types, support for color-management, accessibility features, and much more.

Solution: A facility that promotes retirement of old software and drives adoption of common practices in handling PDF features that developers choose not to support.

5.3 Problem 3: PDF is feature-rich, but not all vendors want to be

Many PDF features are optional. For example, relatively few vendors as yet support digital signature or 3D features in PDF.

When a vendor chooses to support a given feature, it should do so as fully and correctly as the specification requires, and do no harm (whenever possible) to unsupported features. It should warn the user if harm is unavoidable. Today, however, some software fails to warn the user that it will destroy a part of their document!

Solution: A practical and potent means of promoting best practice in creating and processing PDF files. This solution is essentially the same as that identified in section 5.2.

5.4 Problem 4: Fonts need not be embedded for conformance with the specification

Unembedded fonts (permitted but usually inadvisable in PDF) are perhaps the single largest source of unrecoverable problems users encounter. Even in 2014, font problems are not unusual [11]. Although most modern software embeds font subsets by default, font programs remain some of PDF's most complex substructures. Mangled font encoding or a missing ToUnicode entry, for example, is not uncommon.

Solution: Recovering PDF files with missing or damaged font information (among other fatal errors) is sometimes possible. When it is not, providing definitive information about the error and supporting free, high-quality, interactive font substitution would mitigate support costs and enhance vendors' relationships with end-users and digital preservation professionals alike.

5.5 Problem 5: No model for validation

The PDF specification lacks a concept of validity. Neither PDF 1.4 nor ISO 32000 offers much guidance for getting it right, so "does it work in Adobe's Reader" became the fundamental real-world test for non-Adobe software developers (and Adobe's as well, for that matter).

In addition, PDF has a variety of subset specifications. It can be difficult to be sure which specification a file should be validated against, and how. For example, PDF/UA-1 requires the Scope attribute for standard structure type <TH>, but Scope was defined in PDF 1.5. Can a PDF/UA-1 file conform to PDF/A-1a, which is based on PDF 1.4? How do we get a ruling on that question?

It is possible to validate for PDF/A-1b conformance. The specifications for PDF's archival subset standard require specific resources and prohibit certain features. Even so, PDF/A is not obvious in certain cases, and itself relies on the PDF specification. The PDF Association's 2008 Isartor Test Suite [4], was a collaborative effort to resolve many of these problems for PDF/A-1b. Since publication, Isartor has garnered substantial acceptance well beyond the original participating vendors.

Solution: A canonical model for PDF validation would provide a framework for solving all the solvable problems related to PDF reliability and utility in both business and archival contexts. Archivists are aware of this possibility [15]. How do we get there?

6. CANONICAL PDF

The PDF Association has begun studying a concept tentatively named *VeraPDF* [9]. In the next section this paper discusses the concept, and what it could mean for digital preservationists.

6.1 If validators disagree, do they exist?

In the early days of PDF/A collisions between validators were not uncommon [14], which opened fundamental questions about their value. Ensuing customer disappointment prompted development of the Isartor Test Suite, which helped smooth disagreements between different software packages and enabled PDF/A's undeniable success in the marketplace.

Although it is possible that Isartor could be, in general terms, a model for validation of ISO 32000, the prospect is daunting. Isartor would be hard to scale. In itself it does little to promote implementation, and the Terms of Use prohibit using it to certify software products. It is not a solution for canonical validation.

Intended from the outset to serve as a canonical reference implementation, VeraPDF would address the need directly.

6.2 Why "canonical" matters

As mentioned in the introduction, "canonical" validation means a definitive (accepted industry-wide) understanding of compliance with the specification. Knowing that a given feature is implemented in a canonically valid manner it becomes possible to precisely assess the degree of accuracy and completeness with which a given piece of software creates or processes the feature.

In order to simplify matters for those presently concerned only with accurate rendering, for example, it might be argued that conformance with the formal specification is less important than attaining some relative, needs-specific measure of acceptability.

Such an approach, however, offers an unstable, unreliable target. A file may be acceptable in one viewer or when processed through one tool, but not acceptable in another, often as a function of features employed on specific files. This is not a recipe for reliable high-volume processing or long-term preservation.

The problem is especially acute when considering PDF features beyond basic rendering of text and graphics objects. For example, Apple's Preview may in most or all cases render PDF page content as accurately as Adobe's Reader, but as of April 2014, Preview ignores PDF/A, digital signatures and tagged PDF, and even destroys these features when saving a file [5].

A canonical approach sets clear performance expectations. In this context, even when they choose not to fully process a given feature, developers have concrete, impartial guidance at-hand. They are more likely to handle real-world PDF files in a consistent fashion. Open source and industry-accepted file-format validation is how we get there.

6.3 The PDF Reference, in action

VeraPDF would be an open source generic PDF parser similar to EpubCheck [3]. VeraPDF would process the entirety of PDF-defined structures and utilize extension mechanisms to facilitate processing of objects defined elsewhere: font programs, images, JavaScript and other features PDF files may include.

Architecture is always critical, but especially for a purpose-built, future-proofed validator. Ideally, VeraPDF would facilitate modules implemented in both Java and C++ environments and in various programming languages or using 3rd party protocols, and integrate unit-testing resources.

Error handling would allow processing deep into poorly constructed PDF files. Programmatically accessible and localizable reporting for developers would be complemented by industry-accepted "plain language" messages for end-users.

It is important to emphasize that generating useful results from real-world files is not a trivial task because PDF includes such a rich set of features and PDF files may be broken is so many creative ways. It will take an industry effort, but canonical validation offers substantial value to software developers from accelerated software development and reduced support costs.

VeraPDF libraries would be deployable from creation to curation across the entire document lifecycle. VeraPDF could operate as a service or integrate into PDF creation and processing applications including the ingest components of digital repository software.

Beyond establishing the parser's scope and framework the likely initial implementation objective would be validation of classical cross-reference tables, integrating selected grammars such as Adobe's Dictionary Validation Agent (DVA) plugin as potential sources for validation of primary PDF structures. The software can then evolve to meet feature-requests, cover distinct use-cases, highlight best practices, advise on optimization, and more.

One can readily imagine a fantastic open-source validator that understands every aspect of PDF and provides every desirable facility to developers who wish to contribute extensions for non-PDF objects found in PDF files. And yet, such software, if it existed, would not itself answer the key questions:

- How do we know it is canonical?
- What will drive its adoption?

6.4 What makes it canonical

Similar to other infrastructure technologies like plumbing or WiFi, a specific PDF validation model becomes canonical when the industry agrees to treat it as such. There is little question that developers would love canonical quality assurance (QA) tools. If and when the specification's remaining ambiguities and validator policy questions are resolved, and the software developed, then:

- PDF vendors will use it to distinguish conforming from non-conforming software, eventually displacing older or poorly-executed products from the market.
- End-users will use it to evaluate their software and understand (and hopefully, fix) their non-conforming files.

6.5 Adoption drivers

Solving the problems discussed above will require investment by both PDF software developers and those focused on ensuring long-term access to electronic data. The industry collaborations facilitated by the PDF Association's Competence Centers such as the Isartor Test Suite and the Matterhorn Protocol [13], show that for PDF, validation models can thrive in an industry-wide context.

Is VeraPDF achievable? The core value proposition of PDF is interoperability, and the PDF industry knows it. Recognizing the need, the EU created the PREFORMA project [21] to fund development of a purpose-built open source PDF/A implementation checker together with an institutional policy checker. PREFORMA's explicit objective is to become a generally adopted reference implementation.

Hosting the VeraPDF engine on a publically-accessible webserver akin to the W3C's HTML validator [22] with an appropriate interface could provide the functionality indicated in Table 1.

Table 1. Objectives for a canonical PDF validation service

Problem	The VeraPDF Public Validator
1. PDF is complex	Canonical developer education using language accepted by the vendor community
2. Bad PDF software	Collects bad files, identifies the software producer and provides definitive problem identification and corrective information. When possible the server also fixes the file
3. Incomplete support	Drives adoption of best practice via warnings and advisories
4. Problems with fonts	As with Problem 2, provides a mechanism for pooling corrective information
5. No model for validation	Provides developer-centric features to accelerate development and reduce support costs as well as delivering authoritative 3rd party conformance information to end users

6.6 Is canonical validation realistic?

Beyond their protean nature, PDF documents may include a rich mixture of complex, variegated features. It might thus be argued that developing an open-source canonical validator is unrealistic due to the effort required. Adobe has doubtless invested hundreds of man-years in the Adobe Reader, so why would development of a validator be any less daunting? There are three basic reasons:

- A substantial proportion of Adobe's development effort is focused on handling and fixing corrupt or malformed PDF files. Although a useful validator must be able to parse deeply into corrupted files, it need only report its findings.
- Adobe's efforts must meet diverse end user needs and deliver an end user UI and attractive features in a myriad of contexts. By contrast, a validator is a purpose-built developer tool with minimal UI requirements.
- Although the required development effort certainly exceeds the resources readily available to the preservation community, as previously noted, a truly canonical validator has strong appeal to the commercial software world. Such a project will not depend on preservation community resources at all; commercial software interests can drive it.

6.7 How the preservation community can help

The development of a canonical PDF validator will not be trivial, either as a technical matter or in terms of mustering the required collaboration. Since industry acceptance is critical, adoption of the project is most likely to succeed if it is industry-led. The digital preservation community can help make it happen in several ways:

- **Ask for it.** The new NARA Transfer Guidance requires file formats be "valid" according to the format's specification. Encourage procurement entities to require specific assurances from vendors as to the validity of their output.
- **Lobby for it.** The PDF software space is broad and deep, ranging from Microsoft, Google, Apple and Adobe to one-developer shops. Digital preservation professionals know many of the people who develop software and set policy in these vendor organizations. Let them know your priorities.
- **Be a part of it.** From code contributions (for example, to the PREFORMA project) to discussion forums to writing informative error messages and serving on management or policy boards there will be a variety of ways for developers and preservation policy experts to join the effort.

7. CONCLUSION

As ISO 32000, PDF is openly and democratically managed; a *de facto* public trust. Reliability is the bottom line for PDF (and even more so for PDF/A), but ISO committees cannot write software.

While PDF is undeniably the best currently-available format for fixed-form self-contained documents, it is not yet as reliable as it should be. Developers, authors, consumers and archivists alike will all benefit from a concept of valid PDF. Working with commercial software developers the digital preservation community can take a leading role in helping to move PDF from the best available option to the ideal format for now and forever.

8. REFERENCES

[1] Arms, C., Chalfant, D., DeVorsey, K., Dietrich, C., Fleischhauer, C., Lazorchak, B., Morrissey, S., Murray, K. The benefits and risks of the PDF/A-3 file format for archival institutions. NDSA Standards and Practices Working Group 2014-02-20. Retrieved 2014-02-28 from the Library of Congress: http://www.digitalpreservation.gov/ndsa/working_groups/documents/NDSA_PDF_A3_report_final022014.pdf

[2] *De facto* standard, Wikipedia. Retrieved 2014-04-04 from Wikipedia: https://en.wikipedia.org/wiki/De_facto_standard#Examples

[3] EpubCheck, IDPF. Retrieved 2014-04-04 from Github: https://github.com/IDPF/epubcheck

[4] Isartor Test Suite, PDF Association 2011-08-03. Retrieved 2014-03-29 from PDF Association: http://www.pdfa.org/2011/08/isartor-test-suite/

[5] Johnson, D. Apple's Preview: Still not safe for work, Duff Johnson Strategy & Communications 2014-04-07. Retrieved 2014-07-21 from Duff Johnson Strategy & Communications: http://duff-johnson.com/2014/04/07/apples-preview-still-not-safe-for-work/

[6] Johnson, D. The 8 most popular document formats on the web, Duff Johnson Strategy & Communications 2014-02-17. Retrieved 2014-03-26 from Duff Johnson Strategy & Communications:

http://duff-johnson.com/2014/02/17/the-8-most-popular-document-formats-on-the-web/

[7] Johnson, D. Interest in PDF vs. other formats, Duff Johnson Strategy & Communications. Retrieved 2014-03-26 from Duff Johnson Strategy & Communications: http://duff-johnson.com/articles/interest-in-pdf-vs-other-formats/

[8] Johnson, D. 98% of .com is HTML, but 38% of .gov is PDF, Duff Johnson Strategy and Communications 2014-03-10. Retrieved 2014-03-26 from Duff Johnson Strategy & Communications: http://duff-johnson.com/2014/03/10/98-percent-of-dot-com-is-html-but-38-percent-of-dot-gov-is-pdf/

[9] Johnson, D. PDF Validation, Dream or Yawn?, Duff Johnson Strategy & Communications. Retrieved 2014-03-29 from Duff Johnson Strategy & Communications: http://duff-johnson.com/wp-content/uploads/2014/01/PDFValidationDreamOrYawn.pdf

[10] King, J., Introduction to the Insides of PDF, Adobe Systems 2005-04-26. Retrieved 2014-03-29 from Adobe Systems: http://www.adobe.com/content/dam/Adobe/en/technology/pdfs/PDF_Day_A_Look_Inside.pdf

[11] Knijff, J. van der. Identification of PDF preservation risks: analysis of Govdocs selected corpus, Open Planets Foundation 2014-01-27. Retrieved 2014-03-20 from Open Planets Foundation: http://www.openplanetsfoundation.org/blogs/2014-01-27-identification-pdf-preservation-risks-analysis-govdocs-selected-corpus

[12] Leurs, L. The history of PDF, Prepressure.com 2013-08-09. Retrieved 2014-03-27 from Prepressure.com: http://www.prepressure.com/pdf/basics/history

[13] Matterhorn Protocol, PDF Association 2014-02-11. Retrieved 2014-03-29 from PDF Association: http://www.pdfa.org/publication/the-matterhorn-protocol-1/

[14] Moore, R., and Evans, T. Preserving the Grey Literature Explosion: PDF/A and the Digital Archive. Information Standards Quarterly, Fall 2013, 25(3): 20-27 http://dx.doi.org/10.3789/isqv25no3.2013.04

[15] Morrissey, S., "The Network is the Format: PDF and the Long-term Use of Digital Content", *Archiving 2012*, (2012): pp. 200-203. Retrieved 2014-03-23 from Portico: http://www.portico.org/digital-preservation/wp-content/uploads/2012/11/TheNetworkIsTheFormat.pdf

[16] NARA Transfer Guidance, National Archives and Records Administration. Retrieved 2014-04-03 from NARA: http://www.archives.gov/records-mgmt/policy/transfer-guidance.html

[17] PDF (Portable Document Format) Family, National Digital Information Infrastructure and Preservation Program 2014-02-08. Retrieved 2014-03-22 from the Library of Congress: http://www.digitalpreservation.gov/formats/fdd/fdd000030.shtml

[18] PDF/A Competence Center, Nearly all archives projects use PDF/A, PDF/A Competence Center 2009-03-31. Retrieved 2014-04-04 from the PDF Association: http://www.pdfa.org/2009/03/nearly-all-archiving-projects-use-pdfa/

[19] Rimkus, K., Padilla, T., Popp, T. and Martin, G. Digital Preservation File Format Policies of ARL Member Libraries: An Analysis. *D-Lib Magazine*, 20 (3/4): http://dx.doi.org/doi:10.1045/march2014-rimkus

[20] Rosenthol, L., ISO 32000 Document management Portable document format PDF 1.7, Inside PDF Blog 2008-01-28. Retrieved 2014-03-22 from Adobe Systems: http://blogs.adobe.com/insidepdf/2008/01

[21] Tender, PREFORMA. Retrieved 2014-08-03 from PREFORMA Project: http://www.preforma-project.eu/tender.html

[22] W3C Markup Validation Service, W3C. Retrieved 2014-03-29 from W3C: http://validator.w3.org/

A next generation technical registry: moving practice forward

Peter McKinney
National Library of New Zealand
Cnr Molesworth & Aitken St
Wellington
New Zealand
peter.mckinney@dia.govt.nz

Steve Knight
National Library of New Zealand
Cnr Molesworth & Aitken St
Wellington
New Zealand
steve.knight@dia.govt.nz

Jay Gattuso
National Library of New Zealand
Cnr Molesworth & Aitken St
Wellington
New Zealand
jay.gattuso@dia.govt.nz

David Pearson
National Library of Australia
Parkes Place,
Canberra ACT 2600
Australia
dapearso@nla.gov.au

Libor Coufal
National Library of Australia
Parkes Place,
Canberra ACT 2600
Australia
lcoufal@nla.gov.au

Kevin DeVorsey
National Archives and Records Administration
#1 Bowling Green
Room 450
New York, NY 10004
United States of America
Kevin.DeVorsey@nara.gov

David Anderson
University of Portsmouth:
Future Proof Computing Group
Eldon Building
Winston Churchill Avenue
Portsmouth PO1 2DJ
david.anderson@port.ac.uk

Janet Delve
University of Portsmouth:
Future Proof Computing Group
Eldon Building
Winston Churchill Avenue
Portsmouth PO1 2DJ
janet.delve@port.ac.uk

Ross Spencer
Archives New Zealand
10 Mulgrave Street, Thorndon
Wellington 6011
New Zealand
ross.spencer@dia.govt.nz

Jan Hutař
Archives New Zealand
10 Mulgrave Street, Thorndon
Wellington 6011
New Zealand
jan.hutar@dia.govt.nz

ABSTRACT

In this paper we introduce the work of the National and State Libraries Australasia Digital Preservation Technical Registry project.

Digital preservation practitioners must be able to assume technical and intellectual control of content they are charged with preserving. Our experiences tell us that the information and services used to underpin this control are insufficient. Enterprise-class digital preservation services require something better. We believe the solution outlined here is well placed to deliver information required to preserve digital content. Ultimately, this means that the practitioner can say with a strong degree of certainty that they do indeed have control of the content they are charged with preserving.

General Terms

Infrastructure, communities, strategic environment, preservation strategies and workflows, specialist content types, digital preservation marketplace.

Keywords

Technical registry, formats, hardware, carrier media, operating systems, community, NSLA.

1. INTRODUCTION

The digital preservation practitioner, working within the constraints of their institution's mandate has to be able to assume physical and intellectual control of digital objects and maintain that control for the long-term. Physical control requires them to be able to store the file and protect it from harm and further, understand any risks that may relate to its encoding. The nature of that storage and protection is dependent on the mandate and preferences stated at the national, professional, institutional and personal level.

Practitioners are not immediately (if at all) concerned with the actual content of the file or the context of its creation: who the author is, the purpose the record was created, or story told in the book, or the historical importance of the audio. They are fundamentally concerned though with intellectual control through a technical understanding of the file. Principle questions to be answered as they undertake their work include:

- Can I retrieve this file from the medium it is on?
- What format is this in?
- Can I render this file?
- What are the key details of this format that might impact rendering?
- How long will I be able to render it for?
- Should I consider a undertaking a preservation action?
- What can I use to undertake preservation actions on this content?
- What are other practitioners' experiences?

Our experiences tell us that answering these questions with the current tools and services available, while not impossible, requires that results be gathered from many unconnected sources, which can be questionable in terms of their veracity. In general, these results are pitched at a level that is acceptable only for a high-level technical understanding of a file or format.

Missing from this current landscape of tools and information resources is a holistic view of all strands of technical information required to preserve digital content. In addition, where information is available it is often sporadic and incomplete.

Enterprise-class digital preservation services require something better.

In July 2012, the Chief Executives of the National and State Libraries of Australasia (NSLA) approved funding to investigate developing a Digital Preservation Technical Registry (DPTR). This work is undertaken under the auspices of the Digital Preservation Working Group of NSLA.[1] In order to ensure the project captured the best available thinking in the Registry space the NSLA led project team was assembled with a mix of NSLA and international expertise. The project team comprised: the National Library of New Zealand Te Puna Mātauranga o Aotearoa (NLNZ), National Library of Australia (NLA), the National Archives and Records Administration (NARA) in the United States, the University of Portsmouth (UoP) and Archives New Zealand Te Rua Mahara o te Kāwanatanga (ANZ).[2]

The aim is to develop and sustain a Technical Registry (the Registry) that will be a repository of core technical and relationship information for the file formats, computer applications, hardware and media that have been used to encode (and can be used to decode for human consumption) the digital objects that make up digital collections around the world. This comprehensive, consolidated information resource will be able to be used in conjunction with any digital preservation repository in order to support institutions in their efforts to preserve the digital objects in their care.

2. Problem space

In an effort to extend the traditional concepts of physical and intellectual control to digital collections, digital preservation programmes strive to understand how the digital objects in their collection are encoded. They should know what file format each object is encoded in, as well as the format's technical characteristics, dependencies and requirements. Formats evolve through time and as a result often change dramatically, while their names and external identifiers (for example a PRONOM PUID) often remain unchanged across versions. Additionally, application developers often misinterpret specifications or intentionally vary from their instructions, resulting in digital objects that may require special attention. A registry must endure as a resource of reliable, accurate and comprehensive information capable of describing the variations that are known. This information may be stored locally by individual institutions but, due to the complexity and scope of this domain, we are convinced that it will be more efficient to store this data in a collaboratively designed, developed and maintained registry. It will include descriptions of technical environments and the perceived risks to each whether individually or in combination. That is; file formats, software applications, media, hardware, operating systems and input/output devices.

Over the last few decades there has been activity in the form of collaborative discussion (via wikis, other on-line fora, formal conferences, hackathons, and other workshops) and research to identify information, define and validate models, tools, methods, and other mechanisms that are needed for long-term preservation of digital content. To date, much of this work fits the profile associated with "hobbyist" and "artisan" epochs [5]. There is an increasingly urgent need to move to an "industrial" model capable of supporting enterprise-class digital preservation programmes.

We do not believe that previous or current efforts fully meet the needs of a robust, scalable, enterprise-class digital preservation programme. Consequently, there is a lack of a global, consolidated, open, flexible, authoritative, and trustworthy registry of technical information. There are various impacts on the digital preservation community including the time and effort required to find, interpret and match the necessary information from dispersed sources and the potential to undertake work based on insufficient, erroneous or out-dated information.

This project is intended to extend previous work (whether local or global) including PRONOM[3], the Unified Digital Format Registry (UDFR)[4], Mediapedia[5], TOTEM[6], the Planets Core Registry[7], Just Solve It[8], and the current expressions of technical information used in the Rosetta[9] and Safety Deposit Box[10] systems, which are

[1] http://www.nsla.org.au/projects/digital-preservation

[2] http://natlib.govt.nz/, http://www.nla.gov.au/, http://www.archives.gov/, http://www.port.ac.uk/, http://www.archives.govt.nz.

[3] http://www.nationalarchives.gov.uk/PRONOM/Default.aspx.

[4] http://udfr.cdlib.org/.

[5] https://www.nla.gov.au/mediapedia.

[6] http://keep-totem.co.uk/.

[7] http://www.openplanetsfoundation.org/planets-core-registry.

[8] http://fileformats.archiveteam.org/wiki/Main_Page.

[9] http://www.exlibrisgroup.com/category/RosettaOverview.

based on the PRONOM model. Work began in November 2012 to create a vision and logical data model for the proposed Registry in line with the following assumptions.

1. A technical registry supporting preservation risk management, planning and action is central to an ongoing active digital preservation programme.
2. It is undesirable that there should be a multitude of incomplete technical registries globally.
3. A successful registry will have a clearly defined and understandable data model that will enhance user understanding of the data it holds and allow them to make informed decisions.
4. A successful technical registry should be able to provide data to digital preservation repository systems (e.g. Rosetta, SDB, FEDORA, DuraSpace, Archivematica, RODA etc.).
5. A successful technical registry should be more effective than individual products or services that would be required to maintain an active digital preservation programme, e.g., NLNZ Metadata Extractor, JHOVE, DROID and FITS.

2.1 Current Situation

2.1.1 International strategic imperatives

The international digital preservation community is now at a stage of maturity that is a step beyond the advocacy and awareness raising that was a feature of activities at the beginning of the century. National bodies exist, organisations have experience in operating some level of preservation systems as business-as-usual and first-generation tools and services have been developed. This maturity has allowed the community to begin to assess the status quo and lay down some priorities and strategic markers for movement to the next stage of digital preservation activity.

The National Digital Stewardship Alliance (NDSA) in the United States brings together over 160 organisations who wish to advance the practices of preserving digital resources. The NDSA has recently launched an Agenda to highlight gaps and areas requiring development in digital preservation within the United States. The *National Agenda for Digital Stewardship* [9] contains a number of priorities that the Registry would help support. These include "File Format Action Plan Development", "Integration of Digital Forensics Tools" and "Preservation at Scale". The Registry will provide information and services that will directly support these three priorities.

In Britain, the Digital Preservation Coalition (DPC) works from its *DPC Strategic Plan 2012-2015* [10]. As primarily an advocacy body, the DPC does not directly undertake preservation work, but it has objectives to facilitate "knowledge exchange" and "partnership and sustainability" [10, p1]. The Registry, as a community resource and hub will support the DPC members requirements around digital preservation and the DPC itself could play an important role in the sustainable model of the Registry.

The DPC also commissioned the *Mind the Gap* report. This states that "All organisations need to encourage an international 'market' for digital preservation tools by linking up with other projects around the world and engaging with software vendors. This would deliver economies of scale and reduce risk for individual institutions" [11, p7]. In addition, "[o]rganisations should consider the long-term preservation characteristics of the formats they use." [11, p7] The Registry should be the key resource for both of these activities. The registry will ultimately be home to tools used by the digital preservation community; the centrality of the Registry benefitting their ongoing development and fitness for purpose. It will also be the central resource for risk analysis information about formats and actions to mitigate those risks.

UNESCO convened a meeting of experts in 2011 and developed a declaration on digitisation and preservation [12]. This declaration argues that "digital preservation should be a development priority, and investments in infrastructure are essential to ensure trustworthiness of preserved digital records as well as their long-term accessibility and usability" [12, p2]. It also calls on the UNESCO Secretariat to: "establish a multi-stakeholder forum for the discussion of standardization in digitization and digital preservation practices, including the establishment of digital format registries"[12, p2].

It is clear that there is strong alignment of this proposal for a Digital Preservation Technical Registry to NSLA, National and International priorities and strategic directions. Through:

- supporting the preservation and access of content for the benefit of all citizens;
- the supply of trusted information for digital preservation programmes that will engender trust in their activities and the content they preserve;
- supporting a community that will promote collaboration, develop best practices and peer review Registry information.

Two of the strongest imperatives running through the strategies, policies and agendas mentioned are those of trust and collaboration. The Registry supports both of these goals. Through the supply of comprehensive high-quality, peer-reviewed information, organisations can demonstrate that the actions taken are based on best practice thus reinforce or otherwise improve the trust placed in its custodianship of digital materials. At the heart of the Registry will be a community of practitioners and organisations committed to the long-term preservation of digital content. This community will co-create new information, review existing information and help develop tools to take advantage of the information in the Registry. This community will also share their experiences and allow the collaborative creation of best practice. We also hope that the development of the Registry will be a collaborative exercise with various partners including digital preservation organisations and private sector vendors.

2.1.2 Current technical information

As has been stated above, the five member organisations of the project team posit that the current state of technical information for digital preservation is insufficient.

The concerns can be split into two groups. The first set of cover issues with separate information sources. From the format world alone:

- sources vary in terms of the breadth of information they contain (PRONOM holds records on over 1,000 formats, but the Library of Congress around 350);
- sources vary in terms of the depth of information they contain (TRiD contains a very small amount of

[10] http://www.digital-preservation.com/.

information for every format record, but PRONOM has the capability to record a large amount of information);

- there is little (accessible) historical view of technical information. Is Format A still Format A as I understood it five years ago? [4].

The second set cover issues with the entire information space.

- Information sources rarely reference each other.
- Information sources do not agree on how to describe the world (what *is* a format?)
- There is no central community resource that links technical information with community discussion.

These are not strawmen created for the purposes of supporting this project. These concerns impact the partners' directly as they undertake their business-as-usual practices to preserve the records and/or documentary heritage of Australia, the United States and New Zealand. They have also been borne out by the results of a community dialogue exercise. We have presented our work, including our view of the problem space to a number of organisations either undertaking digital preservation research or actively pursuing a digital preservation programme.[11] Every organisation agreed that the current information landscape is not fit for purpose and limits preservation capabilities. Not one organisation said that the status quo was acceptable.

3. The Proposed Solution

The Digital Preservation Technical Registry (the registry henceforth) will do five key things:

1. bring together technical information sources into a central resource;
2. generate new content and relationships that cover a large percentage (i.e. 80-90%) of content existing in collections;
3. allow users to create new content;
4. allow users to build relationships across all information contained in the Registry;
5. allow the community to comment, discuss and share findings on or related to information contained in the Registry.

In order to make these capabilities, the underpinning data model had to take into account existing information sources and offer a change in direction for some aspects of technical information.

3.1 Model

Each of the project team's institutions had existing data models and/or requirements that formed the basis of the logical data model developed. The model is based therefore on TOTEM for hardware and software[12], Mediapedia for carrier mediums[13] and the internal work of NLA, NARA, ANZ and NLNZ [2, 3, 4] in the format area.

The logical data model developed contains five key entities (as shown in Figure 1).

- Hardware
  Information about the mother board, RAM, CPU and Storage. It also includes devices which support the functioning of a computer like data ports, a computer mouse and removable storage devices.
- IO Device
  Information about auxiliary devices such as a keyboard or hard drive that connects to and works with the computer in some way. Other examples of IO Devices are expansion cards, graphic cards, microphones.
- Software
  Information about applications, operating systems and libraries that can be used to create, edit, render, migrate or emulate files.
- Carrier Medium
  Information about the type of medium upon which data may reside.
- Format
  A "particular arrangement of data or characters in a record, instruction, word, etc., in a form that can be processed or stored by a computer" (Oxford University Press, 1989).

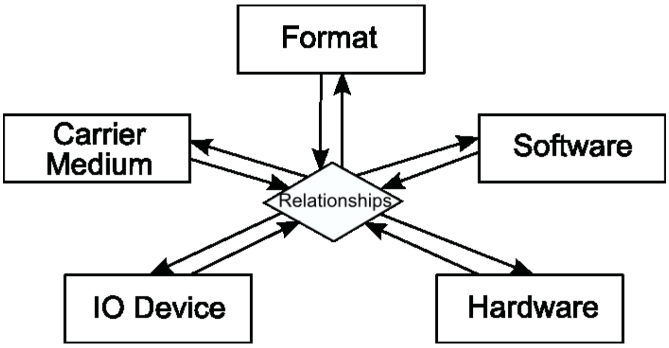


Figure 1: High-level Conceptual Model

While the carrier, software, IO and Hardware aspects of the model are based on existing data models, the format model has been totally re-imagined. It uses three classes of format: Specification, Implementation and Composition. These model the ways in which digital preservation practitioners interact with formats and content

[11] Participating organisations included National Libraries, large collecting institutions and organisations with funding and national strategy mandates.

[12] See http://www.keep-totem.co.uk/.

[13] See http://www.nla.gov.au/mediapedia.

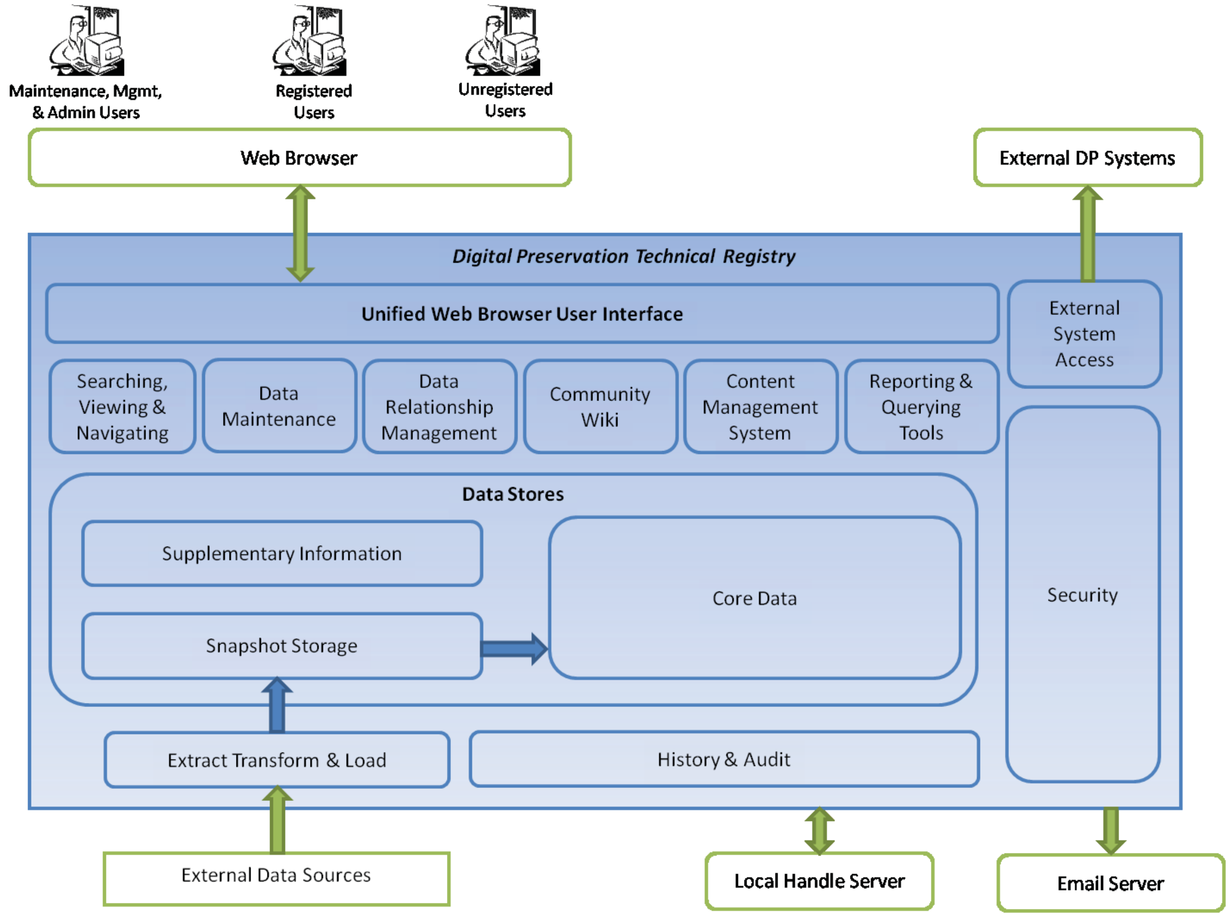


Figure 2: Functional Composition of the Registry

that is represented in those formats.[14] A critical component of the new format model is the concept of an "Aspect". These are the properties that comprise the format types, they are the discrete features and characteristics that are used to build varieties of formats.

The heart of the Registry is the relationships between the entities. It allows all the separate types of information to come alive and become meaningful.

3.2 Functional view

Figure 2 takes a functional composition view of the Registry.

The Registry will give the digital preservation community the following capabilities.

- Ability to import information from current and potential future source registries.
- Ability to store past versions of the external source registry records.
- Ability to support internal registries and online maintenance of the internal registries.
- Ability to flexibly link records within and across external source and internal registries.
- Ability to define the valid link types that can exist between records.
- A web-based user interface.
- Ability to configure what a user, role, or institution can view by allowing information to be filtered based on these attributes.
- Support for creating and running reports across external source and internal registries.
- An API available for external system data export.
- An architecture that supports a decommissioned external source registry becoming an internal supported registry.

[14] The format work is described in more detail in a forthcoming paper.

4. What does this mean?

For the digital preservation practitioner, it means that a whole cosmos of information is available to them and that it resides in one place. It will offer them a breadth and depth of information that is currently unavailable.

Clearly, as can be inferred from the above, the Registry will contain large volumes of information. One way of visualising the information in the Registry and how users will be able to comprehend all the information can be to use the analogy of the night sky. Every piece of hardware, software and media information, every aspect of every format are stars, planets, moons, comets and asteroids.

A wide variety of people ‘interact’ with the night sky. The more experienced the night-sky-watcher, the more detailed their knowledge and more depth they engage with. Large objects are easily identifiable to anyone: a child can see and identify the moon and milky way. As experience of the sky watcher grows, constellations (relationships enforced upon the sky by man) can be identified and used as tools.

At the far end of the scale of experience, the professional astronomer uses high-powered telescopes based on earth or in space to grapple with the universe. These experts use different modes of retrieving information (x-ray, ultraviolet and broad-spectrum views) to understand space from different angles and analyse things that cannot be ‘seen’.

The experience and requirements of the digital preservation practitioner will impact on the level they interact with the information in the Registry. They can stay at the highest level of description and identification (“this is a TIFF”) or can delve through the layers of information and begin to grapple with this cosmos of technical information. They can break down that TIFF file into a version, reflect on the properties (aspects) that comprise it, understand how they impact rendering or preservation activities and converse with other experts on those properties.

Likewise they can understand that they have just a 3M-Scotch magnetic tape. Or they can go deeper and understand that it was created under product code 139, rather than product code 140.[15]

The deeper the interaction with the information, the more meaningful the information. Once the practitioner has knowledge of the exact type of magnetic tape they have, they can understand the impacts of having content stored on that exact variety. Once they know the exact type of TIFF they have (and the exact properties) they can ensure that they are making rendering or preservation decisions based on the best information available. This depth also makes community interactions more meaningful. The question “why won’t this PDF validate in JHOVE” suddenly becomes “why won’t this PDF with encryption and key-length of 128 (Registry ID=xxx) validate in JHOVE 10.2b (Registry ID=yyy)?”

The power of this depth of information is clear. The Registry allows for persistent identifiers to be assigned to such levels of understanding. Users can therefore identify the content they have and bind their relationships and community conversations to that level. It should be noted, that systems or institutions that use existing resources (such as PRONOM) will still be able to use and reference those sources. The Registry will allow for full referencing of those sources and also have the added benefit of allowing users to have historical views of those sources (something that is currently not possible).

Ultimately, this means that the practitioner can say with a strong degree of certainty that they do indeed have intellectual control of the content they are charged with preserving.

At a higher-level, the Registry has the potential to bring a number of benefits to the digital preservation community.

- Trustworthy, high quality information
- More granular understanding of digital collections
- Supporting collection management
- Increased trust in activities
- Efficiency gains
- Economies of scale
- Shared experiences and knowledge
- DP tools utilise Registry

A technical registry is a fundamental component of digital preservation. By moving the current state of the art forward the entire practice of digital preservation benefits.

5. Next steps

Our current work is focused on generating enough collaborative interest in order to build the Registry. A business case has been developed. This proposes a preferred option of international collaboration supporting the build of the Registry and the transition to business as usual. It is clear that the hardest part of the work is not the modeling or requirements capture, nor indeed the build. Rather, the most challenging part will be the transition to a business-as-usual service. The business case therefore focuses not only how to achieve the build, but the transition from completion of the build to a sustainable business.

If successful, this would be a resource built collaboratively and sustained by the community (including the vendors operating in the market). This will require that the digital preservation community consider the weaknesses of the resources currently available, determine how such services can be improved, and ultimately decide the responsibilities of community member institutions to invest in and support a registry that will be of benefit to all.

6. ACKNOWLEDGMENTS

We acknowledge the funding and support from the National and State Libraries of Australasia (NSLA). We also acknowledge the in-kind contributions from each participating organisation. Bill Ross and Dilip Cheerala from Equinox Consulting provided modeling and requirements-capture expertise.

7. REFERENCES

[1] Delve J, & Anderson, D. 2013. *The Trustworthy Online Technical Environments Metadata Database – TOTEM*. Hamburg: Verlag Dr. Kovač.

[2] Gattuso, J. 2012a. *National Library of New Zealand-DROID, PRONOM Developments at the National Library of*

[15] In this case the base material (polyester versus acetate) is different. [http://mediapedia.nla.gov.au/browserecord.php?-action=browse&-recid=110; & http://mediapedia.nla.gov.au/browserecord.php?-action=browse&-recid=111].

New Zealand. Paper presented at Preservation and Archiving Special Interest Group (PASIG), Dublin. Retrieved from http://lib.stanford.edu/files/pasig-oct2012/04-Gattuso_PASIG_presentation_2012.pdf

[3] Gattuso, J. 2012b. *Throughput efficiencies and misidentification risks in DROID*. Retrieved from http://ndha-wiki.natlib.govt.nz/ndha/attach/ReadingResources/MSB%2BDROID%20v1_05.pdf

[4] Gattuso, J. 2012c. *Evaluating the historical persistence of DROID asserted PUIDs. R*etrieved from http://ndha-wiki.natlib.govt.nz/ndha/attach/ReadingResources/Historical%20View%20of%20format%20via%20DROIDv4_2.pdf

[5] McKinney, P., *et al.* 2012. *From Hobbyist to Industrialist. Challenging the DP Community. Paper presented at iPRES 2012, Retrieved from* http://digitalpreservationchallenges.files.wordpress.com/2012/09/mckinney.pdf

[6] Oxford University Press. 1989. *Oxford English Dictionary, 2nd ed.* 3 December 2013. Retrieved from http://www.oed.com.

[7] UC Curation Centre. 2012. *Unified Digital Format Registry (UDFR) Final Report.* Retrieved from http://udfr.org/project/UDFR-final-report.pdf

[8] Webb, C., Pearson, D., & Koerbin, P. 2013. ”Oh, you wanted us to preserve that?!” Statements of Preservation Intent for the National Library of Australia’s Digital Collections. *D-Lib Magazine*. January/February 2013, 19:12.

[9] National Digital Stewardship Alliance. 2013. *National Agenda for Digital Stewardship,* http://libraries.ucsd.edu/news/_files/2013/ndsa-natl-agenda-cover-2014.pdf. Accessed 9 January 2014.

[10] Digital Preservation Coalition. 2011. *Our digital memory accessible tomorrow. DPC Strategic Plan 2012-2015,* December 2011. http://www.dpconline.org/component/docman/doc_download/713-dpcstrategicplan2012-15. Accessed 9 January 2014.

[11] Digital Preservation Coalition. 2006., *Mind the gap. Assessing digital preservation in the UK,* 2006. http://www.dpconline.org/index.php?option=com_docman&task=doc_download&gid=340. Accessed 9 January 2014.

[12] UNESCO/UBC Vancouver Declaration. 2012.. *The Memory of the World in the Digital Age: Digitization and Preservation, 2012,* http://www.unesco.org/new/fileadmin/MULTIMEDIA/HQ/CI/CI/pdf/mow/unesco_ubc_vancouver_declaration_en.pdf. Accessed 9 January

[13] Del Pozo, N., Long, A. S. and Pearson, D. 2010. ‘“Land of the lost”: A discussion of what can be preserved through digital preservation’, in *Library Hi Tech* Vol.28, No.2, pp.290-300.

[14] Pearson, D. and Webb, C. 2008. ‘Defining File Format Obsolescence: A Risky Journey’, *The International Journal of Digital Curation* (IJDC), Issue 1, Volume 3 (July 2008), pp.89-106. Retrieved from http://www.ijdc.net/ijdc/article/view/76/78.

Developing Costing Models for Emulation Based Access in Scientific Libraries

Euan Cochrane
Yale University Library
Digital Preservation Department
New Haven, Connecticut, USA
euan.cochrane@yale.edu

Dirk von Suchodoletz, Klaus Rechert
Faculty of Engineering
Albert-Ludwigs University Freiburg
Freiburg i. B., Germany
{givenname.surname}@rz.uni-freiburg.de

ABSTRACT

Digital preservation practitioners are beginning to answer questions about the costs related to the long-term availability of digital information. In order to successfully model these costs archival systems and workflows need to be fully understood and their costs identified. This can become exceedingly difficult for complex access and preservation strategies like emulation. If emulation is to be considered in a strategy mix its cost components need to be gathered and understood so that institutions can develop informed preservation plans and decide which strategy to follow. The digital preservation community now has a systematic understanding of storage and repository administration costs, but emulation and surrounding services are still an emerging topic for memory institutions. While costs to produce bit-preservable representations of digital artefacts are relatively well known there is an array of rather unpredictable cost factors that need to be further researched. Many of these unexplored costs factors vary depending on the kind of digital objects and the objectives of the stakeholders involved in the activities.

General Terms

Case Studies and Best Practice

Keywords

Cost Model, Emulation, EaaS, Digital Object Access, Archival Workflow, Service

1. INTRODUCTION

Libraries, archives and museums already hold a substantial quantity of digital artefacts and receive an increasing number of digital-born objects with more and more complex structures. These objects require different handling from traditional analogue and static material. Complex digital artefacts must undergo new treatment with regard to methods and workflows to render them accessible to future users which requires memory institutions to implement or acquire from third-party novel types of services.

From a cost perspective digital preservation can be modeled as any other economic activity, i.e. as a usage-based service, or alternatively, the costs of digital preservation services can be designed following the approach used for insurance services. Incentives exist for funding digital preservation services when the benefits outweigh the costs of participating. The advantages of preserving digital artefacts extend from the fact that the discoveries of the future rely on the work of the past. Additionally, for research data, the maintenance of a complete and accurate scholarly record is essential for continued progress in research and learning [6].[1] Cost and business models for emulation services can be derived from a variety of different perspectives. Associated costs can vary heavily depending on object classes and levels of inter-institutional cooperation. Preservation planning and different levels of acceptable risk also influences costs as well as future stakeholders' expectations [23, 9].

Costs have a significant influence on the choice of a preservation strategy, but are inherently hard to quantify. Ultimately, the Total Cost of Ownership (TCO) can be the guiding figure for deciding whether or not a preservation strategy meets the needs of an institution within the constraints of its budget [10]. In addition, there is growing demand for understanding the costs of emulation services within memory institutions and further afield.[2] Institutions looking to implement emulation solutions are currently ill-equipped to do so, partly because there is little information available to provide their funding bodies on how much it might cost to do so. The TCO is also very useful for informing acquisition decisions for collecting institutions. Something that may appear to be a good-value acquisition that is well within the budget of an institution may turn out to be a cost-drain on the organisation once the total cost of ownership is taken into account. For these reasons, and in order to choose appropriate long-term preservation strate-

[1] E.g. to fulfill the requirement for reproducible code in computational science `http://www.recomputation.org/blog/2013/04/12/the-recomputation-manifesto/`.

[2] See various news articles on use of emulation to rescue old hardware, e.g. [15, 11].

gies, and assess preservation plans, proper cost models for using emulation solutions need to be available.

This paper focuses on costs which are directly and indirectly related to institutional emulation strategies. It takes the institutional perspective of a library or archive and ignores traditional repository and bit-level storage costs as they are already several noteworthy articles available on that topic [2, 10, 4, 23].

2. RELATED WORK

A consensus exist that the cost of preservation action must not exceed the estimated value of the digital object [8]. Nevertheless, it might be not entirely clear how to evaluate values of digital objects in different domains [12, 9, 19]. Economic models can be distinguished from cost models and business models, each of which is useful and may be essential for understanding an economic process, but neither of which can be used reliably except in the context of a broader economic model [9].

Early cost models for preserving digital information projected traditional library operation into the digital realm, estimating the efforts required to run repositories and access systems for documents like electronic volumes [22]. The model assumed that all equipment and software costs were capitalized over a life of five years and then replaced for obsolescence. The same cycle was projected for media refresh because of technological change, copying the objects from one bit-level storage to a new one. Equipment maintenance and operations costs were calculated as a proportion of the original purchase price. The personnel costs generated by management and systems engineering services were estimated as a proportion of the salary of a full-time employee including inflation.

The LIFE[2] report discusses possible preservation costing aims and approaches. To cost digital preservation activity two ways have been identified: A top-down audit of all preservation and repository activity; and a bottom-up lifecycle costing of activities relating to a particular content stream [3].

The JISC commissioned the development of application-neutral cost models for digital research data including consideration of different data collection levels and their requirements, the need for relevant documentation and metadata [6, 7]. One of the core goals of "Keep research data safe" was to identify potential sources of cost information. Recommendations hint that institutions repositories should take advantage of economies of scale, using multi-institutional collaboration and outsourcing as appropriate. Typically, once core capacity is in place additional content can be added at increasing levels of efficiency and lower cost.[3] The EU-sponsored 4C project[4] tries to boost uptake of the tools and methods that have been developed. The main objective of 4C is not to develop just another cost model but to ensure that where existing work is relevant, stakeholders realise and understand how to employ those resources.

Successful digital preservation requires long-term planning. There is growing demand for "paid-up" cost models for digital preservation services[5] in order to be able to include provision for funding the long-term preservation of digital content produced by projects, within the projects' proposals. Paid-up cost models are also very attractive for institutions who seek to understand the TCO when making acquisition decisions or when deciding whether to accept donations.

To determine upper limits of acceptable costs it can be useful to change perspective: Billing models and use patterns of existing (non-digital) centrally managed repositories are relevant indicators of what content owners can afford to pay for managed storage services – independent of costs and benefits associated with retrieval [10, 1].

3. EMULATION USAGE SCENARIOS

The concept of emulation of legacy platforms has been included in digital preservation discussion for quite a while [20]. Nevertheless, compared to well established tools and workflows for traditional media, the tools and services for emulation like the KEEP emulation framework [16], and Tesella's Safe Deposit Box that was derived from it, or services like bwFLA Emulation-as-a-Service [21, 18, 14], are comparably new and there is not yet a great deal of experience of deploying these tools in memory institutions.

Within institutions working with digital artifacts there at least three primary use-cases for emulation. Emulation solutions can be applied for:

1. Appraising and/or selecting content in difficult-to-access formats or of dynamic, interactive content
2. Normalizing or migrating content between file formats
3. Accessing content and interacting with it

Each of these roles may present quite different usage patterns and therefore may require different cost models to support them. Below each of these scenarios are explored first and that exploration is then followed with an evaluation of the possible cost models that best support them.

3.1 Emulation in appraisal and selection

Emulation is of use when appraising and/or selecting content as it can give users the ability to investigate content within disk images, or within sets of older digital files and open them in software from the era in which they were created. This can give appraisers and selectors a much richer feel for the content they have presented to them and can help provide a much greater level of context than they might otherwise have had available. Emulation also allows all of this work to be undertaken within closed-environments that can be configured to not save any changes that may have been made (inadvertently or otherwise) during the process.

This appraisal/selection use case requires the organisation using the emulation solutions to have access to a limited

[3] See http://www.beagrie.com/KRDS_Factsheet_0711.pdf

[4] See *Collaboration to Clarify the costs of Curation* self-description http://www.4cproject.eu/about-us

[5] See CNI/CDL model https://wiki.ucop.edu/display/Curation/Cost+Modeling/Princeton and http://dspace.princeton.edu/jspui/handle/88435/dsp01w6634361k

set of generic emulatable environments which have multiple software applications installed on them. They might, for instance, require one or two environments for each major operating system with different sets of software installed on each environment. For costing purposes it is useful to note that this scenario involves a limited number of emulated environments used by a limited number of users on a regular basis.

3.2 Emulation for Content Migration or Normalisation

Often the only software that can open a file (or present its contents with full integrity) is the software that created the file or was originally used to open it. This original software can often also save the content of the file into new files with different formats, and even when that is not an option it is normally possible to use operating-system level utilities, such as print-to-file applications, to save content in different, more accessible, formats. This approach can be useful when a memory organisation has a set of files that cannot be opened in modern software but for which the original software is available. Under this ”migration by emulation scenario” content files are opened in original software running on emulated hardware, and the content is saved into a different format that is still accessible in modern software. For costing purposes it is useful to note that this scenario can be broken down into two distinct subsets with different usage patterns:

- **Just-in-case usage** Used for normalising[6] content at point of ingest. This scenario requires on-going access to emulated environments. These environments contain specific applications for each format that the organisation wants to normalise away from. In this scenario usage is unpredictable, and the emulated environments need to be available at all times just in case a file is acquired that requires normalisation. In this scenario the emulated environments are normally used to process only a small number of files at a time.
- **Just-in-time usage** Used for migrating content when software is completely inaccessible. This usage requires access to emulated environments on demand, when needed. The need for the use of emulated environments for just-in-time usage is usually identified well in advance of the actual use of the environments, and normally does not require emulated environments to be available at all times. Usage of emulated environments in this scenario is predictable and they are normally used to process a large number of files at a time.

3.3 Emulation for access

The most common scenario is to use emulated software to access content in old digital files or to interact with dynamic content. This scenario requires an original environment that includes an operating system and application software to be made available via an emulator. That environment is then deployed to access content stored in one or more digital files held by the organisation. This scenario also has multiple usage patterns, including:

- **On-demand use for specific access requests** In this scenario emulated environments are configured and made available via an emulator and/or emulation service on demand. This use-pattern requires software and emulators to be available but does not necessarily require emulatable environments to be pre-configured for immediate provision.
- **Comprehensive use for all content falling into predefined categories** In this scenario emulated environments are deployed to provide access to all artefacts that fall into a certain category (e.g. when the original interaction software is unusable on modern computers). This use-pattern requires pre-configured environments to be available immediately on request, and emulation access services that can scale to meet user-demand.
- **Mixed usage depending on user-community attributes** In this scenario usage would otherwise be the same as in the blanket-use scenario but it is artificially restricted for some purpose leading to low usage. For example access to content may be restricted to reading rooms in the content-controlling institution. This use-pattern would still require pre-configured environments to be available immediately on request but would not require extensive emulation infrastructure that could scale to meet demand.

These use cases identify a number of factors that help to clarify the best approach to provide the necessary emulation solutions:

- Frequency of use of emulation solutions
- Scale of use of emulation solutions
- Uniqueness of needed emulation solutions
- Timeliness required of emulation solutions
- Regularity of usage of emulation solutions
- Data Security requirements

When developing their own cost models organisations need to identify the use cases that are relevant to their institutions and identify the above factors in order to decide how to model, plan for and recoup the costs of providing these solutions.

4. DIFFERENTIATION

There are a number of components that contribute to the cost of using emulation. These cost components differ depending on how emulation is used and in what workflows it is used. Some workflows would supplement existing ones, others are novel. For example, depending on the type of delivery to be used for digital artifacts the artefacts may have to be copied from their original medium in pre-ingest to the

[6] Normalisation usually means unifying various formats of a certain domain like office documents into a single format, which serves as a standard within the receiving institution.

bit-level storage of the memory institution because of media decay and technological obsolescence [13] independent of the chosen long-term access strategy. Studies of media migration were e.g. done by KEEP.[7] The challenges and related processes are well understood and thus not part of the following considerations.

4.1 Emulation cost components

In order to begin developing cost models for providing emulation solutions it is first necessary to identify the source components that contribute to the TCO for the solutions (list of key cost variables and units, [6]). Once these costs have been identified it will then be possible to group the costs into the products and services that may make up the emulation solutions implement in organisations.

There are numerous cost factors that need to be considered when modeling long-term costs for providing emulation solutions. These include:

Costs related to hardware emulation software

- Emulator development, testing and maintenance costs
- Cost to access original hardware to validate emulation accuracy
- Emulator support costs
- Emulator use costs

Costs related to enabling non-expert access to emulators, e.g. via bwFLA Emulation-as-a-Service (EaaS)

- Remote EaaS software development costs
- Remote EaaS software support costs
- Local EaaS software development costs
- Local EaaS software support costs
- Cost to provide EaaS services

Costs related to intellectual property

- Operating system licensing costs
- Software application licensing costs
- Emulator patent-related costs
- Emulator licensing costs
- License management costs
- Software documentation and manuals copyright costs

Costs related to emulator and environment management

- Cost to configure and maintain environments for adhoc immediate usage
- Cost to document environments and provide unique identifiers/handles.

Costs related to documentation and user-support

- Documentation library creation and maintenance
- Cost to provide remote access to
- Cost to digitize documentation
- End user support for obsolete software
- Cost to provide seamless ”on-line” support within emulation solutions

For the purposes of this paper these costs include all staff costs and hardware costs with the exception of costs related to obsolete hardware needed to compare emulators against for quality assurance.

Regardless of the institutional context there are many emulation-related activities that would benefit from collaborative approaches provided as services in order to reduce the costs for each institution. There are many emulation cost components that could be shared across the community including:

- Development and maintenance of emulators
- Development and maintenance of emulation access services
- License management
- Configuration, management and preservation of installed software environments
- A software, file format and hardware documentation library
- Provision of the ability to run emulators at scale

Nevertheless, several non-shareable costs factors remain:

- Licensing
- Running local hardware
- Running emulators at scale
- End-user support at scale

Having identified the various components of cost that contribute to the TCO for emulation solutions it now possible to begin outlining the different ways these costs can be packaged into products and services which can be sold to internal stakeholders and/or clients.

Most emulation solutions and respective costs can be packaged and costed as fixed-cost products or variable-cost services. Table 1 gives examples of emulation related products and equivalent services:

[7] See `http://www.keep-project.eu/ezpub2/index.php?/eng/content/download/19824/99318/file/KEEP_WP1_D1.2a_v4.0.pdf`

Fixed cost ”products”	Variable-cost services
Normalisation/migration environment	Normalisation/migration of ”x” files
Emulatable environment	”x” hours of access to an emulated environment
Emulation software (emulators)	Emulation as a Service
Emulation experts	Emulation support
Software documentation Library	Access to a software documentation library
Software Licence	”x” hours of access to software
Local EaaS implementation/Emulation workbench	Remote access to Emulation as a Service

Table 1: Emulation products and equivalent services

5. POSSIBLE COST MODELS

Having identified the cost components that contribute to the cost of providing emulation products and services, possible products and services that might be used for providing emulation solutions, and scenarios that emulation solutions might be used within it is now possible to outline possible emulation solutions that might be used within organisations and to develop the cost models to support those solutions. Four models relating to four generalised example scenarios are outlined below. These models assume outsourcing the provision of the emulation services and/or acquiring the full solutions from a third-party provider. Costs for doing all of the work in-house would likely differ greatly depending on context, particularly in regards to managing software licensing fees. For example, costs for just running the hardware to support a remote access to emulation service (EaaS) are currently being determined but are definitely much lower than the overall costs included in these example models. The difference in cost is due to the number of factors related to providing these emulation services as a third-party provider, including (but not limited to):

- Administrative costs
- Legal costs
- Marketing/sales costs
- Human resource costs
- Emulator development costs
- Service development costs.

By assuming the provision of these services by a third-party this simplifies the models and helps to enable readers to understand how such services might be accounted for in their organisations. For example, trying to account for all of the cost components that might go into migrating one digital object from one file format to another can otherwise be quite challenging if this was being done ”manually” within an organization. By assuming the provision of such functionality as packaged services the reader is better able to understand how realistic these might be for their organization to implement.

Model 1: small organisation using emulation for appraisal, selection and infrequent access

Considerations: Small budget, no in-house support
Requirements: Access to ”x” emulation environments provided via an intuitive access system for appraisal and sentencing, infrequent access to a diverse set of remotely provided emulated environments for use in interacting with content for end-user access purposes, no automated migration of objects using the service offered
Appropriate Solution: Small comprehensive set of emulation products for appraisal and selection EaaS provided remotely (or locally depending on security considerations) for access purposes
Rationale: In this scenario the organisation requires a comprehensive set of tools to aid in appraisal and sentencing but these tools would be static and could be acquired as products. The organisation has an unpredictable need for emulation tools for accessing its content so would be best to use a service to provide these, especially given the lack of in-house expertise.

Component	Cost/Unit
Number of environments for Selection/Appraisal	15
Cost per environment	$500
Cost of emulation workbench tool	$500
Total cost of Selection/Appraisal emulation products	$8,000
Number of hours of emulation instances in EaaS per year	$ 520
Average cost per hour	$3
Total Cost for EaaS per year	$1,560
Emulation support services per year (including documentation access and end-user support)	$750
Total cost for emulation solution over 5 years	$19,550

Table 2: Example Cost Model 1

Model 2: Medium sized organisation using emulation for appraisal and selection, a medium level of access, and irregular content migration

Considerations: Medium budget, little in-house support
Requirements: Access to ”x” emulation environments provided via an intuitive access system for appraising and sentencing content, access to a limited set of migration-by-emulation environments and services on an irregular basis and access to a large number of environments for accessing its content that would be used for around 5000 hours a year by users
Appropriate Solution: Comprehensive set of emulation products for appraisal and selection, EaaS provided remotely (or locally depending on security considerations) for access purposes, use of migration by emulation services for 1000 files per year
Rationale: In this scenario the organisation requires a comprehensive set of tools to aid in appraisal and sentencing but these tools could be static and could be acquired as products. The organisation has a medium level of need for emulation

tools for accessing its content so would likely still be best off using a service to provide these. The organisation has a limited need for migrating digital artifacts using emulation each year so would likely be best off using a service for these (table 3).

Component	Cost/Unit
Number of environments for selection/appraisal	15
Cost per environment	$500
Cost of emulation workbench tool	$500
Total cost of selection/appraisal emulation products	$8,000
Number of files migrated using emulation each year	1,000
Cost to migrate each file	$0.10
Total migration cost, per year	$100
Number of hours of emulation instances in EaaS per year	5000
Average cost per hour	$3
Total Cost for EaaS per year	$15,000
Total cost for emulation solution over 5 years	$83,500

Table 3: Example Cost Model 2

Model 3: Large organisation using emulation for appraisal and selection and for comprehensive use for content normalisation upon reception of the content

Considerations: Large budget, available in-house support

Requirements: Access to ”x” emulation environments provided via an intuitive access system for appraisal and sentencing, access to a comprehensive set of migration-by emulation environments/services for migrating 150,000 files per year

Appropriate Solution: Comprehensive set of emulation products for appraisal and selection, use of migration by emulation services for 150,000 files per year

Rationale: In this scenario the organisation requires a comprehensive set of tools to aid in appraisal and sentencing but these tools could be static and could be acquired as products. The organisation has an extensive need for migrating digital artifacts using emulation each year. Depending on the variability of the environments needed for undertaking this emulation it might make sense to undertake this using in-house supported tools. If there is extensive variability in needed-environments a services approach might be more appropriate (table 4).

Component	Cost/Unit
Number of environments for selection/appraisal	15
Cost per environment	$500
Cost of emulation workbench tool	$500
Total cost of selection/appraisal emulation products	$8,000
Number of files migrated using emulation each year	150,000
Cost to migrate each file	$0.10
Total migration cost, per year	$15,000
Total cost for emulation solution over 5 years	$83,000

Table 4: Example Cost Model 3

Model 4: Large organisation using emulation for appraisal and selection, as well as for comprehensive access

Considerations: Decent budget, available in-house support

Requirements: Access to ”x” emulation environments provided via an intuitive access system for appraisal and sentencing, access to a comprehensive set of emulation tools for accessing digital artifacts

Appropriate Solution: Comprehensive set of emulation products for appraisal and selection, and access to a large number of environments for accessing its content that would be used for around 100,000 hours a year by users

Rationale: In this scenario the organisation requires a comprehensive set of tools to aid in appraisal and sentencing but these tools could be static and could be acquired as products. The organisation has an extensive need providing comprehensive access to its objects using emulation tools. Depending on the variability of the environments needed for undertaking this emulation it might make sense to undertake this using in-house supported tools. If there is extensive variability in needed-environments a services approach might be more appropriate (table 5).

Component	Cost/Unit
Number of environments for selection/appraisal	15
Cost per environment	$500
Cost of emulation workbench tool	$500
Total cost of selection/appraisal emulation products	$8,000
Number of hours of emulation instances in EaaS per year	100,000
Average cost per hour	$3
Total Cost for EaaS per year	$300,000
Total cost for emulation solution over 5 years	$1,508,000

Table 5: Example Cost Model 4

5.1 Applying example cost models

The cost models outlined above are indicative examples at best. Actual costs for implementing emulation solutions will vary significantly and will depend greatly on the institutional context. For example, if the institution has an extensive legal team on staff then they may be better equipped to deal with the licensing issues. If an institution has emulation experts on staff then they may be able to configure and run some of the services themselves. When developing a cost model for the use of emulation in a particular real-world context an effective approach may be to:

1. Compare the institutional context to the examples out-

lined above and select the model that best fits with the context.

2. Form an initial model based on one of the selected examples.

3. Review the cost components outlined in the previous section to ensure all cost factors have been either: included in a product or service that has been accounted for, or to highlight missing cost components.

4. Add any missing cost-components to the model.

6. PRELIMINARY PRACTICAL RESULTS

A practical access experiment together with the Rhizome project[8] provided insight into dynamic costs of providing the hardware to support this service and possible usage patterns of such a service.[9]

Currently the bwFLA test and demo infrastructure uses older, written off hardware, using 12 machines, each equipped with two physical Intel Xeon CPUs (E5440) featuring four cores each running at 2.83 GHz. All instances are booted diskless (network boot) with the latest bwFLA codebase deployed. Additionally, there is an EaaS gateway running on four cores delegating request and providing a web container framework (JBoss) for the IFrame delivery. To ensure, a decent performance of individual emulation sessions, one emulation session got assigned to a physical CPU core. In total the test setup handled up to 96 parallel sessions.

The bwFLA cluster was evaluated under heavy load after the Rhizome announced access to a certain dynamic object in their collection. The publicity resulted in an overload of the system in a short period and pushed the average usage level to a higher platform. 700 sessions got evaluated, which resulted in an average session time of 15 minutes.[10] Under the assumption of baseline costs of 50 ct/hour for an 8 core machine at e.g. Amazon cloud[11] such a use case would boil down the session costs to about 2 ct/session. These are reasonable costs in such an application. These results can be used as a baseline for evaluation of migration-through-emulation scenarios, as it could be rather well predicted or measured how long a single run takes to complete. These considerations generate a fairly simple cost model for migrations.

7. CONCLUSION

The above example cost models for providing emulation solutions include reference to emulation products and services that do not currently exist or which are in different stages of development. The services, like bwFLA EaaS, still need further development to become really productive. Cost calculations and considerations for emulation strategies are only just beginning to become realistic as products and services are being made available and as memory institutions begin to consider implementing them. Preservation services can be supplied by one institution, or distributed across many. There are decreased marginal costs from sharing efforts and by sharing code-bases and developing open-source tool suites. Additionally, there are decreased marginal costs by cooperatively running a shared infrastructure.

The actual costs heavily depend on the scope of activities in ingest and access. Depending on the depth of analysis and quality assurance of the single object and expectations of future users the amount of manual labour going into it can become excessive and thus difficult to predict. The inherently long-term nature of digital preservation makes service-based cost models an attractive option as it allows for many of the costs to be passed on to those who benefit from them using a just-in-time approach rather than a just-in-case approach.

As discussed, very few of the shareable components are currently available as products or services from third parties (either for or non-profit). Furthermore, many of these shareable costs relate to activities that most organisations most-likely do not have either the money, nor the will to take on alone. These issues highlight a significant gap in the global digital preservation infrastructure that will need to be addressed if emulation based digital preservation strategies are to be successful over the long-term.

A substantial part of the cost-base of repositories consists of skilled staff and these human resources and many existing workflows and practices will not scale appropriately. There will be a need for more automation of processes and metadata generation, software tools for this, and potentially the development of greater collaboration and shared services to lower the entry and operational costs for institutions [5, 17].

8. ACKNOWLEDGMENTS

We would like to thank the Yale University Library for providing input and resources. This work was partially supported by the *bwFLA – Functional Long-Term Access* project[12] and the German Science Foundation (DFG).

9. REFERENCES

[1] Altman, M., Adams, M. O., Crabtree, J., Donakowski, D., Maynard, M., Pienta, A., and Young, C. H., 2009. Digital preservation through archival collaboration: The data preservation alliance for the social sciences. *American Archivist 72*, 1 (2009), 170–184.

[2] Ashley, K., 1999. Digital archive costs: facts and fallacies.

[3] Ayris, P., Davies, R., McLeod, R., Miao, R., Shenton, H., and Wheatley, P., 2008. The life2 final project report.

[4] Baker, M., Shah, M., Rosenthal, D. S. H., Roussopoulos, M., Maniatis, P., Giuli, T., and Bungale, P., Apr. 2006. A fresh look at the reliability of long-term digital storage. *SIGOPS Oper. Syst. Rev. 40*, 4 (Apr. 2006), 221–234.

[5] Beagrie, N., 2008. Digital curation for science, digital libraries, and individuals. *International Journal of*

[8] See `http://rhizome.org/`

[9] See `http://www.openplanetsfoundation.org/blogs/2014-07-09-eaas-action-%E2%80%94-and-short-meltdown-due-friendly-ddos`

[10] This was higher than expected, due to some long running sessions, as most probably the user switched the browser tab and never closed the original EaaS session.

[11] Pricing: `http://aws.amazon.com/ec2/pricing`

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

Digital Curation 1, 1 (2008), 3–16.

[6] Beagrie, N., Chruszcz, J., and Lavoie, B. F. *Keeping research data safe*. HEFCE, 2008.

[7] Beagrie, N., Lavoie, B. F., and Woollard, M. *Keeping research data safe 2*. HEFCE, 2010.

[8] Becker, C., Kulovits, H., Guttenbrunner, M., Strodl, S., Rauber, A., and Hofman, H., 2009. Systematic planning for digital preservation: evaluating potential strategies and building preservation plans. *International Journal on Digital Libraries 10*, 4 (2009), 133–157.

[9] Blue Ribbon Task Force, 2010. Sustainable economics for a digital planet: Ensuring long-term access to digital information. *Final Report of the Blue Ribbon Task Force on Sustainable Digital Preservation and Access* (2010).

[10] Chapman, S., 2006. Counting the costs of digital preservation: is repository storage affordable? *Journal of digital information 4*, 2 (2006).

[11] Cochrane, E., von Suchodoletz, D., and Crouch, M., 2013. Database preservation using emulation – a case study. *Archifacts*, April (2013), 80–95.

[12] Currall, J., McKinney, P., and Johnson, C., 2006. Digital preservation as an albatross. In *Archiving Conference* (2006), vol. 2006, Society for Imaging Science and Technology, pp. 75–78.

[13] Hedstrom, M., 1997. Digital preservation: a time bomb for digital libraries. *Computers and the Humanities 31*, 3 (1997), 189–202.

[14] Liebetraut, T., Rechert, K., Valizada, I., Meier, K., and von Suchodoloetz, D., 2014. Emulation-as-a-Service – The Past in the Cloud. In *7th IEEE International Conference on Cloud Computing (IEEE CLOUD)* (2014), p. to appear.

[15] Loftus, M. J., 2010. The author's desktop. *Emory Magazine 85*, 4 (2010), 22–27.

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

[17] Palaiologk, A. S., Economides, A. A., Tjalsma, H. D., and Sesink, L. B., 2012. An activity-based costing model for long-term preservation and dissemination of digital research data: the case of dans. *International Journal on Digital Libraries 12*, 4 (2012), 195–214.

[18] Rechert, K., Valizada, I., von Suchodoletz, D., and Latocha, J., 2012. bwFLA – A Functional Approach to Digital Preservation. *PIK – Praxis der Informationsverarbeitung und Kommunikation 35*, 4 (2012), 259–267.

[19] Rechert, K., von Suchodoletz, D., Valizada, I., Latocha, J., Cardenas, T. J., and Kulzhabayev, A., 2014. Take Care of Your Belongings Today - Securing Accessibility to Complex Electronic Business Processes. *Electronic Markets - The International Journal on Networked Business 24*, 2 (2014), 125 – 134.

[20] Rothenberg, J., 2000. Preserving authentic digital information. *Authenticity in a digital environment* (2000), 51–68.

[21] von Suchodoletz, D., Rechert, K., and Valizada, I., 2013. Towards emulation-as-a-service – cloud services for versatile digital object access. *International Journal of Digital Curation 8* (2013), 131–142.

[22] Waters, D., and Garrett, J. *Preserving Digital Information. Report of the Task Force on Archiving of Digital Information.* ERIC, 1996.

[23] Wright, R., Miller, A., and Addis, M., 2009. The significance of storage in the ”cost of risk” of digital preservation. *International Journal of Digital Curation 4*, 3 (2009), 104–122.

Making the strange familiar: Bridging boundaries on database preservation projects

Peter Francis
Public Record Office Victoria
99 Shiel St
North Melbourne VIC 3051
+61 3 9348 5645
peter.francis@prov.vic.gov.au

Alan Kong
Public Record Office Victoria
99 Shiel St
North Melbourne VIC 3051
+61 3 9348 5720
alan.kong@prov.vic.gov.au

ABSTRACT

Archive authorities develop information resources to enable public offices to meet their obligations under their jurisdiction's public records laws. Particular care is taken to ensure that these materials equip their audience with the necessary context and knowledge. Our current work with the evaluation of tools and processes for the preservation of relational databases causes us to question whether good documentation will be enough.

In this paper we describe our experiences at the Public Record Office Victoria (PROV), Australia, in developing processes and guidance for the preservation of relational databases. We find that these projects are different to ‘traditional’ transfers, and that their novelty and technical challenges may be made more difficult by organizational and conceptual complexities. We posit that the nature of such projects may require more than the knowledge of what must be done and how it should be done. We reason that these projects may be hindered by the lack of a shared language to communicate across organisational or functional boundaries.

Using database preservation projects as an example, we discuss the potential contribution that theoretical perspectives such as boundary objects (Star), transmission theory (Shannon) and externalization (Norman) may make to our development of guidance and how this may assist the support of cross-functional dialogue. While focused on database preservation projects, this approach may be generalisable to other cross-disciplinary and cross-functional work.

Categories and Subject Descriptors

H.3.7 **[Information Storage and Retrieval]**: Digital Libraries – *user issues*.

General Terms

Management, Documentation, Design, Human Factors, Theory.

Keywords

boundary objects, public records, database preservation, SIARD

Disclaimer: This paper is part of an exploratory research project and as such should not be regarded as endorsed policy by the Public Record Office Victoria.

1. INTRODUCTION

Public records form part of the Victorian jurisdiction’s critical information infrastructure. They embody much of our community’s civic and personal memory. Further, they play an essential role in the legislative and judicial systems, being relied upon as a true account in forensic legal investigations such as Royal Commissions and citizen’s requests under Freedom of Information laws.

1.1 The responsibilities of archiving authorities

The *Public Records Act 1973* (Act) requires that the Keeper of Public Records establish recordkeeping standards for the efficient management of public records.

Underneath these standards is a comprehensive suite of recordkeeping documents including specifications, guidelines and fact sheets, each tailored for a specific audience including records managers, public officers, commercial entities and researchers.

The Act also specifies that the officer in charge of a public office[1] is responsible for carrying out a program of records management in accordance with the standards.

Our focus in this paper is on our role in the production of this guidance.

1.2 'Traditional' records management

The records management function in many public offices will be seen as largely concerned with management of physical records and dedicated electronic document and records management systems (eDRMS). Typically, the records management function is led by the records team within the agency.

The exponential increase of both physical and digital records, combined with the emergence of a number of disruptive technologies, has caused us to reassess the way we develop guidance.

Further, the manner in which information is stored, managed and used has changed dramatically over the years. This has reached a point where no one single unit within an agency could operate in isolation without the expertise and cooperation from other units.

[1] For the precise definition, see: http://www.austlii.edu.au/au/legis/vic/consol_act/pra1973153/s2.html#public_office

1.3 The SIARD Research project

Archive authorities[2] in Australasia have been developing their capacity to archive public records that are stored in non-records management systems, such as business systems. Earlier studies by PROV have resulted in a suite of projects to address this new landscape. One current project, SIARD Research, was commenced to develop our capacity to preserve relational databases from business systems[3]. There are not currently in place the tools or processes to ensure the continuum [14] management of public records in business systems. Trigger events may be when the business system is being decommissioned or otherwise deemed to be at risk.

The SIARD Research project is evaluating the database archiving tool, SIARD[4], for its use in the transfer of public records from business systems to the state archive. In addition to the technical evaluation, we are exploring the end-to-end management processes, the design of our archive infrastructure, and the resource implications of a full-scale program.

This project has led us to consider the similarities and differences presented. For the purposes of this paper, we will discuss those of particular relevance to our topic – those relating to communication and shared understanding.

1.4 Boundary Objects

In their article, *Institutional Ecology, 'translations', and boundary objects: Amateurs, and professionals in Berkeley's Museum of Vertebrate Zoology, 1907-39* [12], Susan Leigh Star and her co-author James Griesemer examined the heterogeneity of scientific work within the Berkeley Museum.

Expanding the interessement model developed by Latour [5] and Callon [1], Star and Griesemer proposed the use of boundary objects as a mediator to engage the diverse actors to obtain and cooperation across multidisciplinary operations, ultimately to achieve a common goal. This model has been widely cited and the concept of boundary objects has been adopted in disciplines including computer science and public policy.

In their 1989 paper, Star and Griesemer [12] identified four types of boundary objects from their case study, although at that time and subsequently [11] made it clear that there were likely to be more. The initial four types were:

1. A repository that is standardized in a manner that allows access by different actors (i.e. a library catalogue)
2. A representation or abstraction that plays the role of an ideal type, serving as a platform to promote cooperation among different actors (i.e. circuit diagram)
3. An object that could be framed in a manner shared by different actors although the content within that object could vary
4. A form that is standardized in a manner that could be used by different actors

In another words, boundary objects could be viewed as a language which is translated and agreed upon, understood and used by two separate yet related actors across disciplines, facilitating them to achieve a common goal [3]. While clearly facilitating the co-ordination of work, however, boundary objects themselves should not be viewed as possessing co-ordinating features [9].

This paper describes some of the communication issues that may be presented by database preservation projects, and our application of a boundary objects perspective to them.

2. COMMUNICATION AND DATABASE PRESERVATION PROJECTS

Database preservation projects indicate a need for considerable use of cross-disciplinary and cross-organisational communication. This may be problematic as mis-communication between parties may introduced inefficiencies or rework into projects. In some cases, it may even contribute to viable projects being deemed unfeasible.

Cross-disciplinary and cross-functional communication problems are not unique to database preservation projects. Many ICT initiatives, for example, must deal with them. ICT projects, however, will generate considerable design documentation – 'as is' and 'to be' models that can be used in discussions with stakeholders. In contrast, our 'project manager' may be the records manager, who may not be widely recognized across the agency. Further, the preservation of databases for transfer to the state archive is unlikely to attract the resources or authority accorded a transformational ICT project, so the budget will not sustain elaborate documentation and the project will not enjoy high visibility. Our task then, is to support these projects within such constraints.

2.1 The draft process

We will first consider a simple process (Figure 1), where we embed the technical processes for database preservation into one that is similar to that used for the transfer of physical records or those from electronic records management systems. In short, PROV provides the standards and guidance for public offices to localise and execute.

The agency (public office) in the model contacts PROV (or accesses our online resources) for guidance on performing the preservation of a database. Armed with these materials, the agency works through the initial preparation (feasibility, planning), the technical preparation, determining the sentencing actions required (what to transfer to archive, what to leave in place, what to delete), the application of the sentencing, conversion to archival format, transfer to PROV, and ingest into our archive.

[2] For the purposes of this paper, archive authorities are bodies charged with responsibilities for the archiving of the public records for a jurisdiction.

[3] We define business systems as information systems that are not specifically designed to support records management. Databases in business systems may contain public records.

[4] Developed by the Swiss Federal Archives, SIARD stands for Software Independent Archiving of Relational Databases. See http://www.bar.admin.ch/dienstleistungen/00823/00825/index.html?lang=en

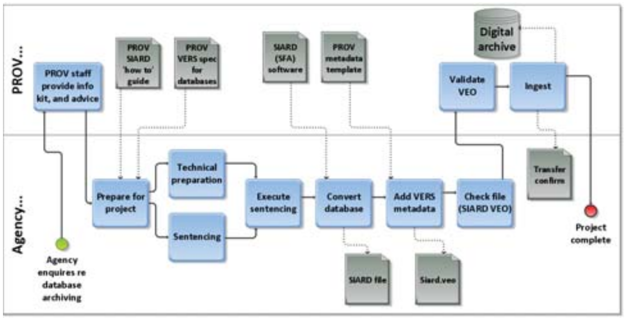


Figure 1: A simple view of the possible process for preserving databases using SIARD.

This is, however, a simplistic view of the process, and one that presumes a homogeneity that is rare in reality. The reality is often more complex, and the progress of such projects made problematic, due to technical, organizational and conceptual complexities not encountered in 'traditional' records management.

2.2 A more realistic view

In reality, accessing, preserving and managing a database within a public office to meet both legislative and organizational requirements will require considerable consultation and collaboration across functional, discipline and organizational boundaries.

2.2.1 Architecturally and technically complex

The tools and techniques for the long term preservation of relational databases continue to improve, however, it remains technically complex when applied to real business systems. The data models may be large and complex, and documentation sparse or non-existent. In the case of older systems, the staff who possess an intimate working knowledge of the system may have moved on.

As depicted in Figure 2, the business system may be accessing data from multiple databases, or the database may be supporting multiple business systems.

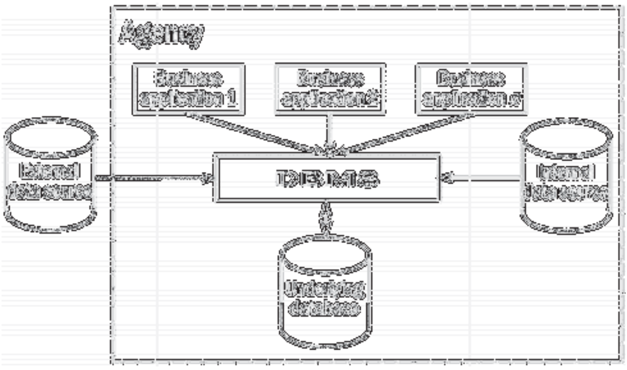


Figure 2: The data forming a 'record' may be aggregated from a number of sources.

2.2.2 Organisationally complex

Modern business systems are rarely managed by a wholly in-house IT function, for example:

- The business system and the database may be managed or hosted by a third party service provider.
- The business system's vendor may maintain control over access to the application layer and the database.
- The business system and the database may be under the management of two different parties.
- The business system or the database may be hosted outside of the jurisdiction.

Gaining access to the database in order to perform the analysis, preparation, preservation and transfer tasks necessary may involve early and ongoing communication among various representatives of the public office (such as records, management and information systems staff), the application vendor, and the IT service provider.

2.2.3 Conceptually complex

The parties who will be involved in a database preservation project will likely bring their own conceptual models and perceptions of the project. As way of illustration, below is a non exhaustive list of the different actors that may have an operational, legislative or contractual interest and responsibility to the same database.

- The **records manager** has an invested interest with the data in the database and will see the database from a records perspective. To a records manager, the primary focus is to ensure that records in the database are preserved, managed, controlled appropriately.
- The **third party service provider** will be contracted to deliver IT services detailed in a suite of service level agreements. This may limit their ability to provide staff or resources to projects, particularly if they are not clearly defined or in terms that could be related to an SLA..
- The **vendor**. The responsibilities of the vendor are usually spelled out in the service agreement with the organization. To a vendor, intellectual property, privacy matters and financial considerations are a priority. They may also perceive general approaches regarding data transfer as an indication that the product is under review.
- The **database administrator**. Someone with database administration duties, and specific knowledge of the source database for the project will need to work on preservation planning and the execution of sentencing and export of the data. They will likely see the database in terms of its data model and stored procedures.
- The **application analyst**. Someone familiar with the business process supported by the system.
- The **data custodian** is someone with responsibility for the data involved to ensure governance obligations are met. In some instances, such as eDRMS may not have data custodians and, if not, this is another group that may have limited exposure to records management.

The participation of many of these people cannot be easily isolated to discrete tasks within the project. In many cases, they may need to work together productively to develop the project from the feasibility stage onwards.

Not only does each actor have their respective discipline based perception on a particular issue, he/she will also have their own psychological biases and work history which will vary even within the same discipline.

Many of these are people who have had little or no prior exposure to the records management environment, which indicates that records management concepts and terms may not be a natural option for a common language.

2.3 Example scenario: sentencing

The sentencing of the records may be done by a records manager, while the execution of the sentencing done by a database administrator or similar.

Records managers need to apply disposal plans to the records in the system (sentencing). To do so, they will need to see the data from a records perspective.

Once sentenced, they will likely need a DBA or similar role to execute that sentencing. The DBA will need the sentencing actions to be in a form that can unambiguously applied to the data model.

When executed, it is prudent that the action be validated - there is a risk here that miscommunications at any point may introduce errors - the wrong data may be preserved/destroyed. That is, neither the records manager, nor the DBA is able to judge that the other's work has been done correctly.

The database and/or the business system may be managed by a third party service provider. This has a number of implications: any work may come at a cost, and that cost may depend upon how 'actionable' the execution plan is (it is better to be in terms that they are familiar with and which do not need reinterpretation). The provider's representative will need to be involved at the feasibility stage - any miscommunication may result in advice that may prove prohibitively expensive making the project unviable, or may result in advice that the project is not technically feasible, or may be interpreted as impinging upon the vendor's IP (which we believe is in fact unlikely in most cases).

Addressing the technical complexities may be made more difficult due to the organizational and conceptual complexities that are likely on database preservation projects.

2.3.1 Addressing the performance gap

We can see that to implement database archiving projects will go beyond existing practices and perceived roles.

- If we expect that the work on SIARD projects will necessarily be across disciplines, as records managers will need to make sentencing decisions but data managers will need to execute them,
- and if records managers and data managers use different concepts and terms, and view the data in different ways,

then we should consider measures to reduce these barriers when developing our guidance materials.

3. BOUNDARY OBJECTS AND DATABASE PRESERVATION PROJECTS

3.1 Relevant qualities of a boundary object

3.1.1 Translation

Further, we believe that the language used and the form of the 'object' must not disenfranchise or subordinate collaborators – there will likely be a leader, but the object should not determine who that will be.

When we think about translations, we do it with Shannon's [10] model in mind (Figure 3), which, although developed for telecommunications, has been found more widely applicable to human communication.

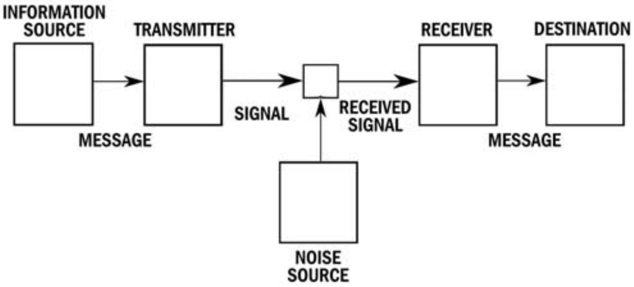


Figure 3: Shannon's schematic diagram of a general communication system [10]**.**

In the non-technologically mediated case of two people speaking to each other, the Transmitter could be regarded as the language and concepts used by the speaker (what they say and how they say it). The Receiver may be the interpretive filter (of their role and experience) that may influence what the listener hears. Although originally a technical model, we find the concept of messages undergoing encoding and decoding helpful. The role of a boundary object may minimize the need for both parties to 'translate' for the other.

In the earlier sentencing example, communication is depending upon the forming of the request by the records manager and the interpretation of the request into database operations by the DBA. Where the need for interpretation, or re-analysis, is high, so too is the risk of error or unnecessary rework.

3.1.2 Externalisation to aid cognition

Although not a strict quality of boundary objects, we anticipate most will have a material quality that will support individual and shared thinking. Externalisations have long been considered to enable memory and computational offloading, freeing the mind of some of the burden during problem solving (see, for example [4, 7]).

We see that a boundary object in database preservation projects that enables a database administrator and records manager to relate the 'record' and the data model to the business system would reduce cognitive load on both parties.

3.1.3 Non-directive and unbiassed

A boundary object is non-directive, it does not embody any responsibilities or agreements, and implies no obligation on the parties. Where such mechanisms are necessary, they can be managed outside of, not through, the object.

The planning model, as demonstrated by Suchman [13], is flawed. We should take care not to build our logic into the object and introduce further barriers to use.

3.2 A boundary object for database preservation

We look for possible common concepts, ones that directly relate to the system, but in which each party can derive meaning for their own work. For example, one candidate that is neither a record nor a data model is the business object.

3.2.1 A business objects perspective of the data

If public records are to be identified and appraised in business systems, it will be necessary to look at the business system's data (a relational database model) from a records perspective. Once records management decisions have been made, they must be translated into requirements that a database administrator can execute.

From a database perspective, Olson [8] describes business objects as either 'entities' or 'transactions'. Entities persist for long periods of time, and are subject to change over time. Transactions are records of events that are created and completed in a relatively short period of time.

3.3 The sentencing scenario revisited

If we consider the case of a fictitious government agency, the Dept of Science. The records manager has identified the Service Delivery System (SDS) as likely holding public records. The SDS supports the department's role in providing advice to research organizations. The Advisory Services function is covered by a Retention & Disposal Authority (RDA), developed by the department to identify their public records and detail their management.

The RDA has been used to manage Advisory Services records stored in the department's electronic records management system, however, the records manager believes that the SDS system contains data that would also be required to be preserved permanently and transferred to PROV.

Figure 4 depicts a simple business object model of the fictional SDS. This view may map well onto the records management concept of a record.

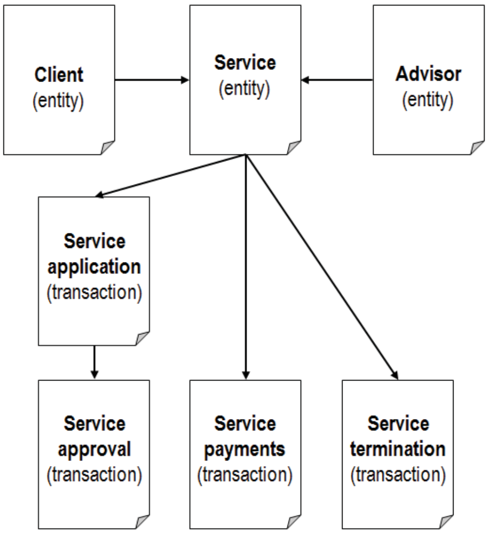

Client (entity)
Service (entity)
Advisor (entity)
Service application (transaction)
Service approval (transaction)
Service payments (transaction)
Service termination (transaction)

Figure 4: An example of a simple business objects perspective as a boundary object (using a fictitious Dept of Science service delivery system).

It may be that by jointly analyzing the business system and expressing it terms of business objects the records manager and database administrator will establish a shared understanding of the system.

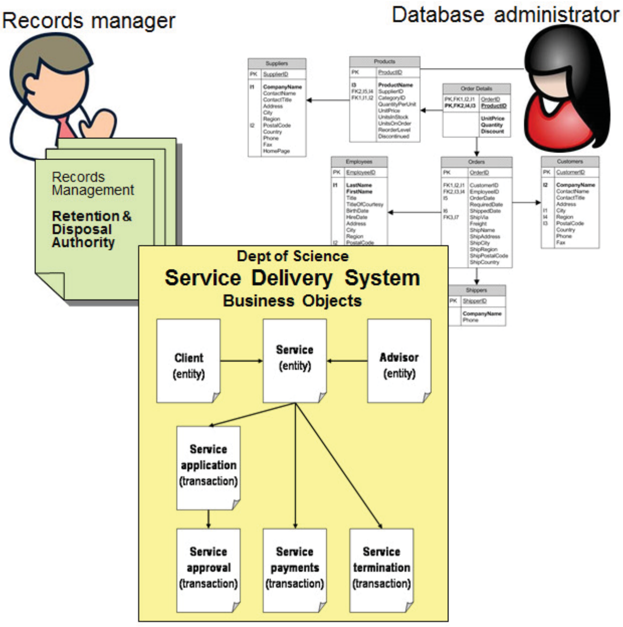

Records manager
Database administrator
Records Management
Retention & Disposal Authority
Dept of Science
Service Delivery System
Business Objects
Client (entity)
Service (entity)
Advisor (entity)
Service application (transaction)
Service approval (transaction)
Service payments (transaction)
Service termination (transaction)

Figure 5: Example of the use of a business objects concept to facilitate communication between a records manager and database administrator.

3.3.1 As a translation support

In the example of use depicted in Figure 5, the business object model may serve as a useful bridge for the records manager to describe the data requiring action, and the criteria for determining action (such as retain, transfer to PROV, destroy, etc.). The RM may find it easier to express the functional descriptions of the RDA into relevant business objects, than on a database schema. For their part, the DBA may be more confident in tracing the database tables and fields supporting a business object, than from the descriptions commonly found in an RDA.

3.3.2 As a form of externalization

By providing a physical model that is able to be expressed as a diagram (as above) a table or list, both the RM and the DBA can reduce the need to retain both the conceptual model and the past determinations as they deal with a problem at hand.

3.3.3 Non-directive and unbiased

The business objects model may be useful to both the RM and the DBA but does not clearly belong to either world. In this way, it does not confer ownership to either.

This exchange highlights another potential benefit in that it may simplify the identification of the data required, in instances where the data is distributed by providing a logical rather than physical perspective.

There are a number of potential barriers that may hinder the adoption of boundary objects. One particular assumption is that each actor, given he/she is fully aware of the type of boundary object that is at play, is willing to adopt the object to achieve an outcome. However, this level of willingness is dependent on a number of factors including the actor's trust of the approach, past history, relationship with the other actor and other behavioral biases.

In addition, the boundary object itself is silent on whether the achieved outcome reflects work policy or the organisation's overall strategic direction. Without addressing these fundamental concerns, it is likely that despite the boundary object being effectively used, there will be no support from the executive or stakeholders.

Boundary objects are unique in that they are designed to address one particular given circumstance which may become ineffective when applied elsewhere.

3.4 Evaluation

We will be using data generated during the SIARD Research project to map records management definitions and concepts onto data models and vice versa. In the process, we will look for opportunities for the development of general principles that can be used as the basis for the development of a transformation tool.

Our initial evaluation of this approach and of any potential boundary objects will be through iterative co-design and collaboration with our project partners. We believe that this field development will give our work a form of member validation [6] and we leave the judgment as to our success to those who it is intended to support.

4. CONCLUSION

The motivation for the work described in this paper is founded on a number of questions: We ask, as we always do, are our guidance materials fit for purpose? Are they accurate? Do they reflect policy? Are they within our scope, not straying into areas beyond our brief? Are they generalisable, do they work for all our public offices?

Our work to date on the SIARD Research project causes us now to ask, will our usual approaches be successful? Is there more than knowing what to do, and how to do it? We must anticipate that database preservation projects will rarely enjoy the resources, design documentation, or profile that would accompany an ICT project. Our proposed approach, outlined in this paper, is shaped by two constraints: the almost infinite variety of installations in public offices, and the clearly finite resources that archiving authorities are able to allocate to any problems.

Business systems and the underlying databases are implemented in a variety of ways and under a variety of management arrangements. Even at the data level, there will be the possible need for operator intervention, and the use of a variety of export and conversion tools. “The processing of the finding aids has taught us many useful lessons relevant to preservation of databases and other structured data. It revealed that there is no such thing as a standard way to import data. Most of the 3.1 million records needed some kind of human intervention during the import process. The data of the DTNA project was imported using a variety of different methods such as direct database connections and exporting data as CSV from the source.” [2] p.9

Archiving authorities cannot always ‘be there’ for the agency, to assist or facilitate – it is not sustainable for them to do so. They can, however, continue to reflect upon the guidance materials they provide.

We have identified that some new approaches to the preservation of public records may be impeded by organizational and conceptual complexities not generally encountered during more traditional public records transfers. The archiving authority may not necessarily be able to address them simply by providing better advice on what should be done, however, including a boundary objects perspective into our thinking as we develop resources to support public offices may assist in better communication and collaboration on cross-disciplinary public records preservation projects.

5. ACKNOWLEDGMENTS

Our thanks to our colleagues at Public Records Office Victoria who worked on this project with us or acted as reviewers for this paper.

6. REFERENCES

[1] Callon, M. 1986. Some elements of a sociology of translation: domestication of the scallops and the fishermen of St Brieuc Bay. *Power, Action and Belief. A New Sociology of Knowledge?* J. Law, ed. Routledge, London. 196–223.

[2] Essen, M. van, Rooij, M. de, Roberts, B. and Dobbelsteen, M. van den 2011. *Database preservation case study: Review*. National Archives of the Netherlands.

[3] Fox, N.J. 2011. Boundary objects, social meanings and the success of new technologies. *Sociology*. 45, 1 (2011), 70–85.

[4] Hutchins, E. 1995. *Cognition in the wild*. MIT Press, Cambridge, Mass.

[5] Latour, B. 1987. *Science in action: How to follow scientists and engineers through society*. Harvard University Press, Cambridge, Mass.

[6] Neuman, W.L. 2003. *Social research methods: Qualitative and quantitative approaches*. Allyn and Bacon, Boston.

[7] Norman, D.A. 1993. *Things that make us smart: Defending human attributes in the age of the machine*. Addison-Wesley Publishing Co., Reading, Mass.

[8] Olson, J.E. 2010. *Database archiving: how to keep lots of data for a very long time*. Morgan Kaufmann.

[9] Schmidt, K. and Bannon, L. 1992. Taking CSCW seriously. *Computer Supported Cooperative Work (CSCW)*. 1, 1-2 (1992), 7–40.

[10] Shannon, C.E. 1948. A mathematical theory of communication. *The Bell System Technical Journal*. 27, (1948), 379–423, 623–656.

[11] Star, S.L. 2010. This is not a boundary object: Reflections on the origin of a concept. *Science, Technology & Human Values*. 35, 5 (2010), 601–617.

[12] Star, S.L. and Griesemer, J.R. 1989. Institutional ecology, 'translations’ and boundary objects: Amateurs and professionals in Berkeley’s Museum of Vertebrate Zoology, 1907-39. *Social studies of science*. 19, 3 (1989), 387–420.

[13] Suchman, L.A. 1987. *Plans and Situated Actions: The Problem of Human-machine Communication*. Cambridge University Press.

[14] Upward, F. 1996. Structuring the records continuum - part one: Postcustodial principles and properties. *Archives and Manuscripts*. 24, 2 (1996).

Automatic Discovery of Preservation Alternatives Supported by Community Maintained Knowledge Bases

Rudolf Mayer, Johannes Binder,
Stephan Strodl
Secure Business Austria
Vienna, Austria

Andreas Rauber
Vienna University of Technology
& Secure Business Austria
Vienna, Austria

ABSTRACT

Preservation Planning, which deals with selecting the most appropriate preservation action to be applied to digital objects, is an important step in any digital preservation activity. Comprehensive Preservation Planning depends on the availability of identified alternatives of preservation actions, which are for example file format migrations to migrate data in an outdated format to one that has better support. Also emulation, e.g. of the behaviour of a specific software application (application emulation), can be a viable preservation action. The alternative identification step can either be performed manually by an expert, or (semi-)automatically, if appropriate knowledge bases are available. Building and maintaining such knowledge bases is however a tedious task, as the number of software applications and file formats, and especially their relation to each other, is very large. In this paper, we therefore present an approach to automatically build knowledge bases for Preservation Planing from already existing, open resources. One such source is the community maintained Freebase, which contains linked data on many topics, among them file formats, software applications, and most importantly, their relations, in a structured manner. We demonstrate the applicability of these knowledge bases by automatically identifying possible digital preservative actions on a uses case, an eScience experiment from the domain of data mining. This use case originates from the task of process preservation, where we look beyond single files, but regard complete chains of executions as the objects to be preserved.

1. INTRODUCTION

Preservation planning can be understood as a form of utility analysis, where each different possible preservation action is quantified. The goal is to select the most appropriate preservation action to be applied to digital objects. Preservation Planning is a vital step in any digital preservation activity.

An important phase in Preservation Planning is to identify viable preservation actions, i.e. to identify which actions can be applied to the digital objects that would prepare them to be usable in the future. Such preservation actions are for example file format migrations to migrate data in an outdated format to one that has better support. In most cases, there is a wealth of possible formats to convert into. Also emulation, e.g. the emulation of the behaviour of a software application, is an important approach in digital preservation.

Business processes are a more complex form of digital objects, where the domain of interest moves beyond single files, but to complete chains of process executions, including a number of files generated and consumed, and the software needed to manipulate them. To provide a faithful preservation of the execution of the process, preserving the behaviour of the software stack required for the process steps becomes necessary. In the setting of process preservation, we thus look beyond single files, but also regard the complete chain of a process execution, and the environment a process is executed in. Therefore, we move from regarding only the view path of single object, towards the more complex interaction of multiple view-paths that might be realised in the same system.

Alternative identification for preservation planning can either be performed manually by an expert, or automatically, if appropriate knowledge bases are available. Building and maintaining such knowledge bases is however a tedious task. In this paper, we therefore present an approach to automatically harvest such knowledge bases for Preservation Planing from already existing resources. Specifically, we utilise the community-maintained Freebase, as well as the domain of Linux software packages. On top of these knowledge bases, we develop a service that can automatically identify preservation action alternatives for a given system. These systems need to be described in a formal way according to a model recently proposed in [2], which introduces a model to describe the context of business processes. As a part of this model, the technical environment of a system can be described.

It has to be noted that the service presented in this paper is meant for the discovery and identification of alternatives. The suitability of these alternatives for actually solving the digital preservation problem at hand still have to be assessed and verified by digital preservation experts.

The remainder of this paper is structured as follows. In Section 2 we give an overview on related work. Section 3 reviews the Context Model, which can be utilised to formally represent the context of a process, of which we are specifically interested in modelling computing systems. In Section 4 we then describe the data sources and harvesting processes to obtain our knowledge bases. Section 5 will then detail on how these knowledge bases can be utilised, in conjunction with the formal mode of a system, to identify preservation alternatives. In Section 6, we then show the applicability of the approach on a use case example. Specifically, we take an example of a process to be preserved, and analyse the different alternatives identifiable. Finally, we provide conclusions and an outlook on future work in Section 7.

2. RELATED WORK

The term *Digital Preservation* as defined in the UNESCO Guidelines for the Preservation of the Digital Heritage [18] is the process of preserving data of digital origin. The two main strategies for the preservation of digital heritage listed are migration [10] and emulation [14, 16, 7]

Emulation refers to the capability of a device or software to replicate the behaviour of a different device or software. Emulation can happen on different levels in a system:

- *Application* An application is usually utilised to render a digital object (if the digital object to be preserved is not itself an application, e.g., computer games, digital art, self-running documents, process management software). By replacing the original application interpreting the digital object the functionality of this application is emulated.
- *Operating System* On a modern computer system an operating system provides access to the underlying hardware for an application running on top of it. By providing a layer that redirects the operating system calls of the application to the same calls of a different operating system, it is possible to emulate the operating system with this additional layer on top of a new operating system.
- *Computer Architecture* The most common use of emulation is to emulate the functionality of a computer architecture by using software, thus introducing an additional layer in the software stack of a rendering environment. Physical hardware can be emulated using either full hardware emulation where all hardware components of the computer architecture are recreated in software on a new host-system or by virtualisation where the CPU is not completely emulated (like in virtualisation software such as VirtualBox[1]).

Regarding emulation, in this work, we are primarily interested in identifying emulation opportunities for applications. However, the model described in Section 5 could also be utilised to identify strategies e.g. for Computer Architecture emulation.

File format migration is a strategy of refreshing digital files over time, to keep the content stored in formats that can be interpreted by current technology. Migration might also be done anticipatory and transform contents to formats that are expected to be readable in the future. Such a migration is usually easier done today, as more tools that can read the presumably outdated format are still available. Identification of suitable format migration paths that are supported by currently available software tools is a primary concern for our approach regarding format migration.

Also for this approach, it is important to have a knowledge base on file formats, and the software that can manipulate it. Several possible sources were investigated, foremost well established registries such as PRONOM , and tools developed in the SCAPE project to facilitate preservation planning. However, these approaches did not provide a comprehensive and up-to-date data base of software that can handle the various formats.

Several attempts to build comprehensive digital preservation related knowledge bases or registries exist. The PRONOM registry[2][5], developed by The National Archives of the United Kingdom, primarily contains information on file formats, along with a classification, description, publication dates, and vendors. Further, the registry provides information on software applications, such versions, release dates, and default file formats for that software. In addition, also vendors are registered. Each entry in the registry's database is assigned a PRONOM Unique identifier. Currently, the registry holds around 1,100 file formats, as well as around 280 entries on software. It also contains basic support for identifying migration pathways, i.e. conversion chains from one format to another, along with the software that supports this. However, the database currently contains less than 50 of these pathways. PRONOM is also designed to contain information on whether a format is at risk, however, this information is generally not provided.

The Community Owned digital Preservation Tool Registry (COPTR)[3] is a registry for tools useful for preserving digital information for the long term. It contains a Wiki-style collection of tools along with a short description of their functionality. However, this information is not well structured, and can't be processed automatically. Also, links to file formats these tools are capable of processing are missing. Currently, the registry contains around 360 tool entries.

While PRONOM and COPTR surely have huge impact on digital preservation solutions that need this type of registry information, it seems that the amount of content provided is not enough for identifying a larger set of alternatives. This was also recognised by [6], where the authors try to aggregate information on file formats from several sources. They utilise linked open data repositories for this approach. We will in the subsequent sections investigate also on some of the sources utilised in that approach.

Comparing different options of preservation actions is the challenge of preservation planning. In [3] a preservation planning workflow that allows for repeatable evaluation of preservation alternatives is described.

[1] VirtualBox – `https://www.virtualbox.org/`

[2] `http://www.nationalarchives.gov.uk/PRONOM`

[3] `http://coptr.digipres.org/`

Regarding the long-term availability of software, the Software Sustainability Institute defines, among others, the following strategies [9]:

- Emulation of the execution environment, i.e. utilising emulators that mimic the functionality and behaviour of the hardware and software environment. This strategy requires *Operating System and Computer Architecture Emulation.*
- Migration of the software to a different platform. This can be as simple as just compiling otherwise platform independent software for the different platform or in worst case may require a complete rewrite of the software.
- Technical preservation of the hardware environment.
- Cultivation, by releasing the software into open source and engage the community to maintain and develop it.
- Hibernation, which includes archiving the software and the knowledge needed to use it, for a potential future use.

Most of these strategies can be useful in the preservation of software applications. However, most of them are rather alternatives that try to preserve the status-quo of the current system setup. They do not require a specific identification step of possible alternatives, which would be the case e.g. for migration of a file format, where we need to know which formats are available for a specific process setup.

The view path [17] of a digital object is the combination of a software and hardware that is required to render an object. This can be described with the *Preservation Layer Model* (PLM), which typically consists of the layers of a specific application, and operating system and the hardware supporting that operating system. However, but more complex layering is possible as well. The above mentioned techniques of migration and emulation basically modify elements in this view path. In the domain of preserving complete processes, which can be understood as a digital object itself, we normally encounter a multitude of digital objects that are manipulated in a chain. Often subsequent steps depend on the output of the previous activity. In such a setting, multiple view-paths exist, and they partly share some of the elements from the different layers, e.g. the same operating system might support two different applications used in two different steps.

3. REPRESENTATION OF SYSTEMS TO BE PRESERVED

A formal model to represent the context a process is embedded in was presented in [2]. In the setting of process preservation, all but the simplest processes require to be described by a multitude of information objects, as well as their interconnections and relations. Examples of the details to be preserved are the process model itself, and the actors involved in the process execution. On a more technical level, the infrastructure required to support the process execution is of interest. This includes the hardware and software that provide the execution platform, as well as various artefacts consumed and created during the process. Of interest are furthermore any dependencies to external parties. To enable a semantic description of these objects in a structured manner, the context model, a formal meta-model, was derived. It describes classes of elements and their relations, in the form of OWL ontologies. To be extensible, it is designed with a core (upper) ontology describing the generic concepts, and extension mechanisms to map supplementary ontologies describing more specific aspects. Ontologies are a well-suited method to implement this architecture.

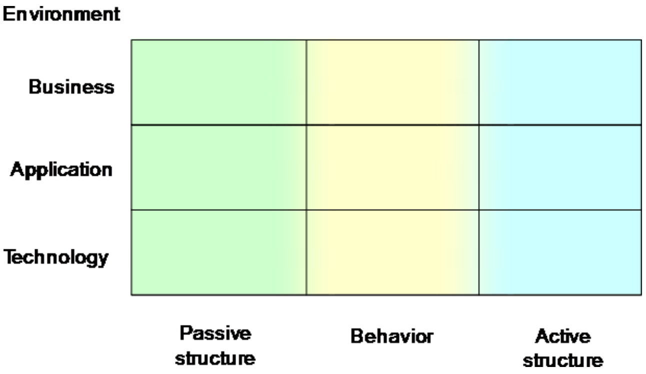


Figure 1: The ArchiMate Framework ([8])

The core ontology is based on the ArchiMate 2.0 language ([8]), an international standard from the Enterprise Architecture domain. The ArchiMate modelling language includes a minimum set of concepts and relationships. The ArchiMate framework organises its language concepts in a 3 × 3 matrix: the rows capture the different enterprise layers *business*, *application*, and *technology*, and the columns capture the cross layer aspects *active structure*, *behaviour* and *passive structure*. Figure 1 depicts this organisation of the framework, while Figure 2 lists the main concepts provided by ArchiMate, where the colours of the elements corresponding to the categorisation into active structure, behaviour and passive structure. Active structure contains entities capable of performing behaviour. The behaviour itself contains elements defined as units of activity performed by one or more active structure elements, and the passive structure contains objects on which the behaviour is performed.

For the task of identifying preservation alternatives, we can use the concepts of the technological layer of the framework to model our systems.

The core domain-independent ontology of the Context Model is then augmented through a set of specific extension ontologies that are tailored to explicit modelling concerns. Currently, the context model provides extensions to cover aspects such as *Legal*, *License*, *Patents*, *Data & Formats*, *Hardware*. The extension ontologies are, when possible, based on already existing languages, for which then the ontology mapping to the core ontology was provided. On overview on this is given in Figure 3. Most of these extensions map to elements in the technological layer, and are thus also of interest for our modelling concerns.

Specifically, the current implementation of the alternative

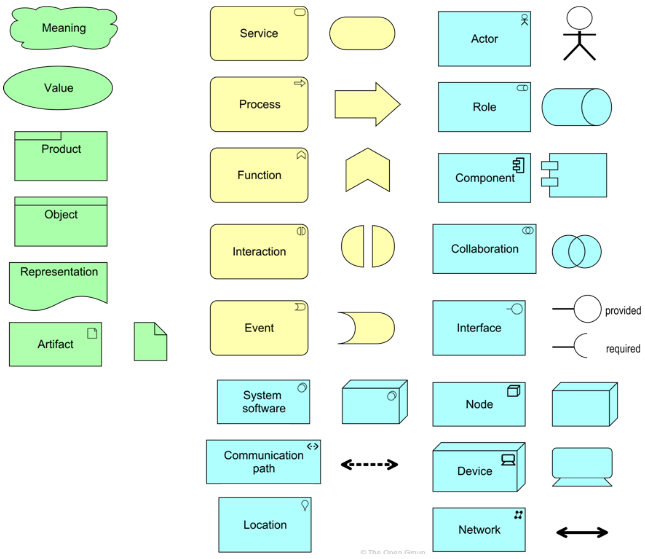


Figure 2: The ArchiMate meta-model

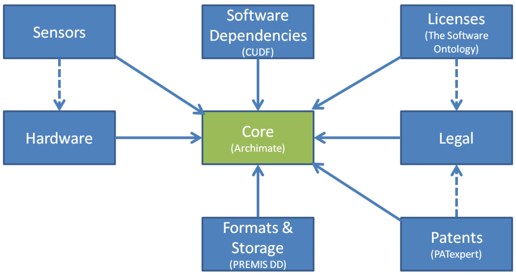

Sensors
Software Dependencies (CUDF)
Licenses (The Software Ontology)
Hardware
Core (Archimate)
Legal
Formats & Storage (PREMIS DD)
Patents (PATexpert)

Figure 3: Overview on available extensions and their relation to the core ontology

identification operates on the following entities: Artifact, SystemSoftware and FileFormat. The two former are part of the core ontology, while the third one is an element defined via the data format extensions, which is realised via the PREMIS data dictionary.

4. KNOWLEDGE BASE GENERATION

In this section, we describe two different approaches to obtain the data needed for the knowledge bases of our alternative identification service. We further discuss technical details of the representation of the knowledge.

4.1 Freebase – Software and File Formats

The online database Freebase[4] [4] provides a community driven and maintained database of semantic linked-data on various topics. Among them, there is information on software applications and file formats. The schema for software tools[5] is described in Table 4.1. Currently, there are more than 9,000 entries in this schema. The schema for file formats[6] is described in Table 4.1. Freebase contains at the moment more than 3,500 entries for file formats.

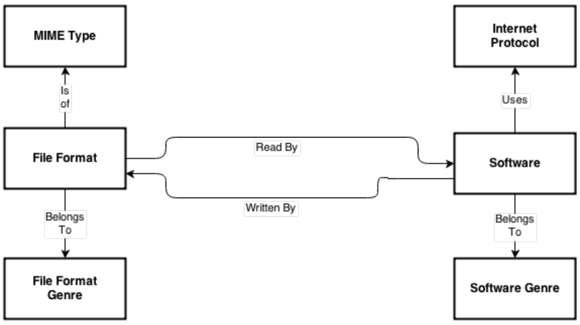


Figure 4: Relations between File Formats and Software in Freebase

An overview of some of the relations in the database is given in Figure 4. The *Written By* and *Read By* properties allow linking software applications to specific file formats. Specifically, this allows on the one hand to identify possible conversion paths from an origin file format to a desired file format, by identifying software tools that can read the origin and write the target format. In more complex cases, if no software is available that can directly do this conversion, chains of format migrations via intermediate formats can be established. On the other hand, the information on which formats can be read by a specific software allows to establish a rudimentary list of software that is compatible to each other. It is possible to deduct which software applications are capable of handling the same types of file formats, and thus, theoretically, exchangeable. Of course this identification of equivalence ignores the functionality provided by each software, and thus might return a list of false-positive equivalents. Also, some of the potential preservation alternatives might not make sense from other points of view. It therefore requires still, as mentioned above, the review and assessment of a digital preservation expert. Another approach of identifying software with similar functionality is via the genre and protocols. The former is a human classification of types, e.g. PDF readers as software that can render PDF files, while the latter can be utilised for software that is no directly manipulating files, such as an FTP client, implementing the File-Transfer Protocol.

While some of the data in Freebase is not as clean as in other registries that are dedicated to digital preservation, it has two rather big advantages. On the one hand, the process of extending the knowledge base is very simple via an online interface, and happens at a frequent rate by the community. Also, due to the linked data scheme, information from Freebase can be easily augmented by other means than directly in the Freebase database, e.g. by augmenting it by a locally available data source. Also the size of the knowledge base is an advantage for the task of alternative identification. At the moment, there are three times as many formats, and 45 times more software applications in Freebase compared to the PRONOM registry.

4.2 Software alternatives for Linux packages

A second approach to build a knowledge base for software application emulation is based on the concept of software

[4] `http://www.freebase.com/`

[5] `http://www.freebase.com/computer/software?schema`

[6] `http://www.freebase.com/computer/file_format?schema`

Table 1: Freebase Data Schema for Software

Property	Description
Developer	Manufactures of the software (e.g. organisation or person)
Software Genre	Categorisation of applications, e.g. *Database management system* or *PDF reader*
First Released	Date of the first release of this software
Latest Version	Version number of the latest release
Latest Release Date	Date of the latest release
License	The license the software is released under, e.g. GNU General Public License
Programming languages used	Programming languages used to write the application, e.g. C++, Objective-C, etc.
Compatible Operating Systems	Name and versions of operating systems the software can be run on
Protocols Used	The Internet Protocols used in this application, e.g. Hypertext Transfer Protocol (HTTP)
Protocols Provider	Other software that also use the same protocols

Table 2: Freebase Data Schema for File Formats (excerpt)

Extension	Common extension of this file format
Genre	Categorisation of formats, e.g. *Audio file format* or *Executable*
Creation Date	The date when the format was created / published
Written By	Link to software applications that can **write** this file format
Read By	Link to software applications that can **read** this file format
Used On	A list of operating system platforms the format is commonly used on
Format Creator	The organisation or individual creating the format
Magic	The magic number (identifier) of this file format, e.g. GIF89a for GIF images
MIME Type	The MIME type of the format
Contained By	The container format this format is usually contained in
Container For	Others formats this format is a container for; e.g., CSO is a container format for compressed ISO images
Extended From	Any other format this format is based on / derived from
Extended To	Any other format that extends on this specific format

packages, used in many Linux distributions, e.g. Debian[7]. In these operating systems, software applications (and components) are normally made available in a specific package format, which is in most cases a specific compressed container format. The package contains the actual software application, as well as control information for the installation process of the package. As such, it provides e.g. scripts that should be run after the software application is extracted to the system, e.g. to perform other changes on the system. One example is the creation of a specific user that would execute a package that provides a server program. Furthermore, control information in the packages provides details on the dependencies of that package. It might e.g. define that for a web server package to be installed, also the Java runtime environment is required. The package manager then automatically handles acquiring and installing also these dependencies.

In these package based operating systems, there is generally a universe of packages that can be installed, and that are known to the package manager. Further, there is the concept of a virtual package, which can be seen as a place-holder package for other (real) packages that then provide the functionality. This concept is also reflected e.g. in CUDF (Common Upgradeability Description Format [15]), which is a format used to describe installation and upgrade paths. Examples of such virtual packages are e.g. “web-browser”, or a “java-runtime“, and a “c-compiler“. These packages then are provided by specific implementations and from the dependency structures defined in the packages, different implementations can be interchanged. For the “java-runtime” package, providers might be OpenJDK[8], Oracle Java[9], or the Cacao Virtual Machine[10]. These packages provide the same functionality according to the Java Virtual Machine specification, but might greatly differ in regards of their implementation and license. One requirement might e.g. be that the used package should have a license that allows obtaining and modifying the source code, to allow modifications in case a changed system environment requires that.

In order to obtain a knowledge base for the software package, we implemented a tool that gathers the virtual packages and their providers for a specific version of distributions of a Linux system. In principle this tool is based on the Debian package system, and thus covers also operating systems based on Debian, such as Ubuntu[11] or Linux Mint[12].

In total, on a current Linux Ubuntu distribution, around

[7] `http://www.debian.org/`

[8] `http://openjdk.java.net/`

[9] `http://www.java.com`

[10] `http://www.cacaojvm.org`

[11] `http://www.ubuntu.com/`

[12] `http://www.linuxmint.com/`

2.000 virtual software packages that have more than one provider can be identified. Not all of these are actual software applications, some are also just components, i.e. virtual packages that are providing libraries that are in turn used in other applications to built end-user applications. Such libraries can e.g. be components for GUI programming, or libraries that allow interfacing with a specific hardware.

4.3 Representation of Knowledge Bases

As a representation format for our knowledge bases, we opted for using ontologies, specifically the Web Ontology Language (OWL) [13], a widely used knowledge representation language. OWL is intended to augment the Resource Description Framework (RDF), and provides formal semantics, as well as RDF/XML-based serialisations. The reason for choosing this representation is that on the one hand, OWL defines several convenient mechanisms to query the knowledge base. Queries can as such be formulated via OWL Description Logic (OWL-DL), or the graph query language SPARQL [1]. Another motivation for choosing OWL ontologies is that the model to represent a system (cf. Section 3), a part of the previously mentioned process context model, itself is authored using the Web Ontology Language. Using OWL for the knowledge bases representation thus simplifies cross-model queries and reasoning.

Freebase provides an API to query the online content. However, we opted to store the data locally for a number of reasons. First of all, the local storage allows for a more efficient querying of the data, as potentially many subsequent queries need to be sent. Furthermore, we also combined the Data from Freebase with information on Formats from PRONOM, by a simple approach of matching along the file extension and MIME Type. Finally, local storage allows us to represent the knowledge base in a form that enables easy automatic reasoning and discovery of migration paths. We therefore developed the ontology that is depicted in Figure 5. The major elements in there are Formats, Tools and Registries. These are further utilised to perform certain actions, such as migration.

For the second knowledge base obtained from the Linux Package manager, we opted to represent this in CUDF (Common Upgradeability Description Format). CUDF is also utilised in the context model presented in Section 3, where it serves as one domain-specific ontology representing package dependencies. A representation of the concepts of CUDF is given Figure 6. The main information entities are a Package and the VirtualPackage; there is a wealth of relations defined, such as depends, conflicts, etc.

In CUDF, virtual packages can be considered to be a kind of categorisation of the concrete packages, similar to the genre provided by Freebase. If we encounter a certain package, we can thus simple query which virtual packages it provides, and then find other packages that provide this virtual package. Alternatively, if the model already uses a virtual package to model explicit what functionality is required, we can query to replace that specific provider.

5. IDENTIFYING PRESERVATION ACTION POSSIBILITIES

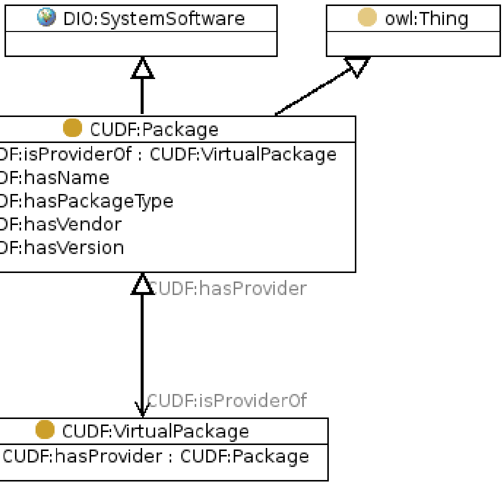


Figure 6: Concepts of CUDF

The alternative identification application currently considers file format migration and software application emulation. We will describe these two in detail below.

5.1 Software Application Emulation

For each software involved in the process (SystemSoftware or Artifact concepts in the Context Model), a software application emulation is proposed, by identifying software that is equivalent to the currently employed applications.

This approach identifies software replacements for a specific software application at risk. Such a risk might be a lack of future support, incompatibility with other components of the process, or that the license the software is published under is prohibitive for the future use or preservation of the system. Using the knowledge base obtained from Freebase, we are able to retrieve migration path information in a structured way. We can e.g. propose the migration of a proprietary word processor file format to a more standard format. Depending on whether we need just read access or also write access to the artefact, different conversions will be available – in general, there will be more support for reading a specific format, thus if this is the only requirement, we will be able to identify more potential alternatives.

The proposed alternative will also take into account which changes in the software stack are needed. To this end, a prototype implementation of a package dependency solver for Linux distributions is being developed, which will be utilised to identify the changes in the software stack. Once a specific file format is identified, consulting the dependency solver will notify us whether the software stack used currently is sufficient to also work with the new file format, or new software needs to be installed. This can in turn mean that a specific software that was previously used to manipulate a digital object is not needed anymore. This may then be removed, and the dependency solver will also be used to determine which other software components that were only needed by the removed application can as well be removed.

5.2 File Format Migration

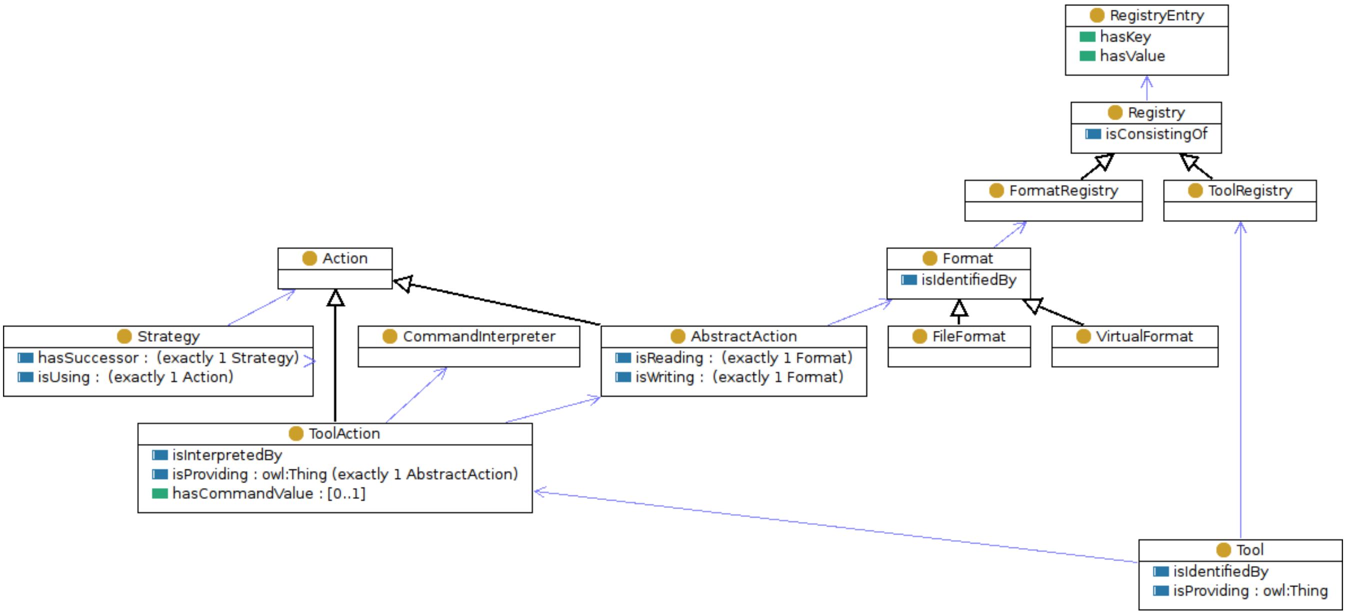


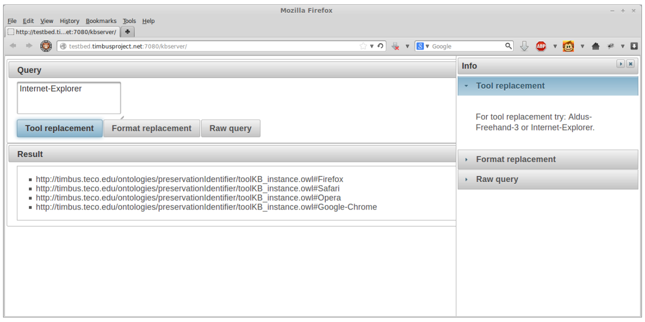
Figure 5: Ontology to store information on migration tools

For each data object (*Artifact* concept in the Context Model) that is either produced or consumed in the process and for which the data format is at risk of becoming obsolete (e.g. a proprietary format for which the vendor support might end), an alternative for ensuring long-term access to this data object has to be produced.

Firstly, once a file format is identified to be at risk and should be replaced, an alternative format providing similar functionality has to be identified. Identifying a similar format can be automated by utilising the *genre* information present in the File Format schema (cf. Section 4), with the straight-forward approach being to identify formats in the same genre, and then select those which are connected via a format migration path using the available software tool migration capabilities.

Secondly, if the new format is not also supported by the software currently available at the system, also the current software setup needs to be modified. In this case, it is need to identify the required changes for the steps in the process that access the files in the old format, and potentially replace the current software applications with different applications that can work with the new format we migrated to. This affects software that reads, writes or renders these files.

Data objects could be both interpreted by humans (in a human processing step in the process, where the human takes decisions based on the content of the data object, or augments/modifies the data object), or by software. In the case of a human task, the exact rendering of the data object e.g. on the screen is important. It is therefore important to select an emulation strategy that preserves this property most faithful. In the case of a machine task, preservation of the data object has to go in hand with ensuring the software can still process the data object, but rendering capabilities are of less importance.



Internet-Explorer

- http://timbus.teco.edu/ontologies/preservationIdentifier/toolKB_instance.owl#Firefox
- http://timbus.teco.edu/ontologies/preservationIdentifier/toolKB_instance.owl#Safari
- http://timbus.teco.edu/ontologies/preservationIdentifier/toolKB_instance.owl#Opera
- http://timbus.teco.edu/ontologies/preservationIdentifier/toolKB_instance.owl#Google-Chrome

For tool replacement try: Aldus-Freehand-3 or Internet-Explorer.

Figure 7: Preservation alternative for replacing "Internet Explorer" by alternative software.

5.3 Online Query interface

The knowledge bases obtained from Freebase and the Linux Package Universe can be queried online for preservation identification alternatives, as seen in Figure 7, which depicts a potential replacement of Internet Explorer by alternative web browser software such as Firefox, Safari, Opera or Google Chrome. In the future, we will also provide an API that could be utilised by other services needing information on file formats and software tools.

5.4 Alternative Identification Output

The output of the alternative identification module is a set of possible preservation action alternatives. These alternatives will then have to be analysed by a digital preservation expert regarding their feasibility and suitability, who will then select those actions that best fit his requirements.

Specifically, each alternative contains a modified version of the technology view of the system, modelled via the process context meta-model, and a list of changes that were done to arrive at this new system, from the original instance. The

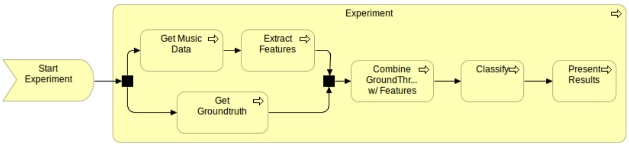


Figure 9: Process model of the eScience experiment

context models, original and modified, are OWL ontologies, as detailed in Section 3. The list of changes is provided by the means of the OWL version of the PREMIS preservation data dictionary. Specifically, the "Event" entity is used to link entities (linkingSourceObject and linkingOutcomeObject), together in a softwareReplacementEvent or formatMigrationEvent. The source and outcome objects are generic PremisEntity elements, of which, via the mapping of the PREMIS extension to the core ontology in the context model, specific software artefacts are instances of. An example of such a change list can in Figure 5.4.

6. USE CASE APPLICATION

The use case we want to investigate in detail is an e-Science experiment in the domain of machine learning. Specifically, it tests the usefulness of a method for automatically classifying items in a music collection into a set of predefined categories corresponding to music genres, by computing the accuracy of the classification (i.e. for how many songs the algorithm can detect the correct genre). It is performed by a researcher which aims to collect performance metrics for classification and make comparisons to the state of the art. The motivation for performing the preservation of such a process is related to any possible challenges to the results that can be made by members of the research community. Thus, by preserving such process, the provenance and authenticity of the results can be proven, and the process can easily be repeated on different data, or with altering the parameters, at a later stage, as well.

A process model is depicted in Figure 9. First, music data and a ground truth ("gold standard") of the genre assignment are acquired from external providers. Then features (numerical representations) are extracted from the music files. These are combined with the gold standard, and converted to a different format, before a classification model is learned. Finally, the performance of that model is evaluated. This process is described in much more detail in [11] and [12].

Figure 10 depicts a graphical representation of the technological infrastructure of the process. Central to the process is the Taverna Workflow engine, in which the process is modelled, and which orchestrates the execution. The audio feature extraction, as well as the format conversion are implemented in Java, and require a version 6 Java runtime to be executed. The machine learning software is provided by the open source Toolkit "Weka" [19], which as well requires Java. Also smaller helper applications to fetch music data and ground-truth are implemented in Java as well. Java is provided in this setup by the Oracle Java 6 implementation, which comes with a restrictive license that disallows redistribution among other things. Important for the process are also the File Formats utilised - on the one hand there is MP3 which is used to encode the audio files, and then there are a series of custom text formats, such as the SOMLib and ARFF Formats. These are defined in respective specification documents, which are authored in HTML and Adobe Acrobat respectively. On the current setup, the closed-source software Safari and Adobe Acrobat are used to view them.

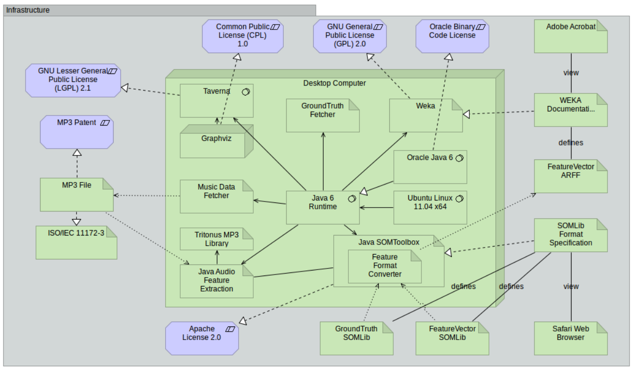


Figure 10: Technical infrastructure model of the eScience experiment

To be able to preserve the technical environment of the process, the following automatic alternatives to the current system can be identified

Software Emulation. Oracle Java is restrictive in regards to source code and redistribution, thus it is preservable to replace the Java runtime implementation by other means. Through the knowledge base on Package Alternatives, we can identify OpenJDK as an open-source alternative. Via the Freebase knowledge base, we can identify similar alternatives. The SOMLib documentation is in the current setup displayed via the Safari Browser. This might not be an ideal candidate for long-term preservation, as no source code is available, and it is thus more difficult to adapt to a changed environment. An automatic proposal would yield e.g. Firefox or the Chromium browser as alternatives. A similar issue arises with the documentation of WEKA, which is in PDF format; the alternative proposals yield the open-source tools Okular and Evince as alternatives.

File Format Migration. The feature extraction service currently takes various input formats, such as WAVE, MP3 or FLAC. In the current experimental setup, MP3 is used, which is processed with the help of a third-party library *tritonus*[13]. This library is however not actively developed since 2003, and frequently has errors with MP3 files that have a slightly unusual encoding. Furthermore, MP3 is partially protected by a patent, and that might cause problems for certain preservation actions to be applied in the future. It might thus be beneficial to change to a different file format. Format replacement would suggest e.g. a conversion to WAVE PCM, using e.g. the software *mpg123*. Of course,

[13] `http://www.tritonus.org/`

```
<ClassAssertion>
  <Class IRI=”http://id.loc.gov/ontologies/premis.rdf#Event”/>
  <NamedIndividual IRI=”[serviceLocation]/[identifier]/SoftwareReplacement”/>
</ClassAssertion>
<ObjectPropertyAssertion>
  <ObjectProperty IRI=”http://id.loc.gov/ontologies/premis.rdf#linkingSourceObject”/>
  <NamedIndividual IRI=”[serviceLocation]/[identifier]/SoftwareReplacement”/>
  <NamedIndividual IRI=”[originalModelURI]#OracleJava1.6”/>
</ObjectPropertyAssertion>
<ObjectPropertyAssertion>
  <ObjectProperty IRI=”http://id.loc.gov/ontologies/premis.rdf#linkingOutcomeObject”/>
  <NamedIndividual IRI=”[serviceLocation]/[identifier]/SoftwareReplacement”/>
  <NamedIndividual IRI=”[serviceLocation]/[modifiedModelURI]#OpenJDK1.6”/>
</ObjectPropertyAssertion>
```

Figure 8: Example of the output describing the changes made to the system by replacing *Oracle Java 1.6* with *OpenJDK 6*

several other tools are available to do this specific migration, such as *lame*, *ffmpeg*, or applications with a graphical user interface such as *mplayer*. In addition, several other target formats are proposed, such as the Free Lossless Audio Codec (FLAC), for which similar conversion tools are applicable.

The current documentation of the SOMLib format in HTML might be risky, as HTML is still an evolving standard (e.g. to currently HTML 5), and it is has shown to be difficult to exactly preserve the behaviour of Documents, especially across different implementations of web browsers. A format migration of HTML would identify PDF as a suitable candidate, using e.g. the tool *wkhtmltopdf*. The software stack also needs to be updated, as we now don't need an HTML viewer anymore.

In Figure 11, we can see one potential candidate modification to the view-paths in the process. We modified specifically viewing the HTML and PDF files, and converted the file format of the music data to WAVE, thus requiring the new application ”mpg123“. Modifications to the specifications were done manually, the licenses can be determined automatically for some software applications, as this information is provided for most Linux Packages, and for some software applications registered in Freebase.

7. CONCLUSIONS AND FUTURE WORK

Knowledge bases on file formats and software applications are an important aspect in digital preservation, especially in the phase of preservation planing, where alternatives of preservation actions need to be identified and evaluated. Existing knowledge bases often lack in the depth and freshness of information provided, as maintaining them is a tedious task. In this paper, we thus presented an approach to obtain knowledge bases on file formats and software applications from repositories such as the community maintained linked open data source Freebase, as well as Linux package repositories. We on the one hand offer these knowledge bases in a publicly available API that can be used by digital preservation solution providers. Further, we also presented a prototypical implementation of a preservation alternative discovery and identification service that leverages these knowledge bases. We demonstrated the usefulness of this approach on a use case evaluation.

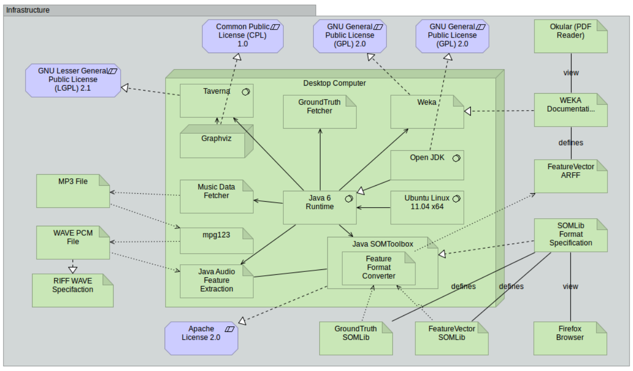


Figure 11: Technical infrastructure model of the eScience experiment, after applying preservation actions to migrate file formats and using alternative software

Future work will focus on fine tuning the software and file format knowledge bases obtained from the online repositories, and improve the alternative identification approach. Potential future extensions to the knowledge base and alternative identification are:

- Emulation of the execution environment, i.e. utilising emulators that mimic the functionality and behaviour of the hardware and software environment (operating system). Information on hardware is available in the context model, thus only information on emulators is mission.
- Migration of the software to a different platform. This can be as simple as just compiling an otherwise platform independent software for the different platform, or in worst case be a complete rewrite of the software. Freebase might offer enough data for this, as information on programming languages utilised for a software, and compilers availability for certain platforms, is available.

Furthermore, we will be applying the alternative identification service to more use cases from the TIMBUS project, among others a use case on monitoring of large civil engineering structures and from the e-Health domain, as well as on other use case from the domain of scientific workflows.

Acknowledgements

This work has been co-funded by COMET K1, FFG - Austrian Research Promotion Agency and by the TIMBUS project, co-funded by the European Union under the 7th Framework Programme (FP7/2007-2013) under grant agreement no. 269940.

8. REFERENCES

[1] SPARQL query language for RDF. Technical report, World Wide Web Consortium, Jan. 2008.

[2] G. Antunes, M. Bakhshandeh, R. Mayer, J. Borbinha, and A. Caetano. Using ontologies for enterprise architecture analysis. In *Proceedings of the 8th Trends in Enterprise Architecture Research Workshop (TEAR 2013), in conjunction with the 17th IEEE International EDOC Conference (EDOC 2013)*, Vancouver, British Columbia, Canada, September 9-13 2013.

[3] C. Becker, H. Kulovits, M. Guttenbrunner, S. Strodl, A. Rauber, and H. Hofman. Systematic planning for digital preservation: Evaluating potential strategies and building preservation plans. *IJDL*, 10(4):133–157, 2009.

[4] K. Bollacker, C. Evans, P. Paritosh, T. Sturge, and J. Taylor. Freebase: A collaboratively created graph database for structuring human knowledge. In *Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data*, SIGMOD '08, pages 1247–1250, New York, NY, USA, 2008. ACM.

[5] T. Brody, L. Carr, J. M. N. Hey, A. Brown, and S. Hitchcock. Pronom-roar: Adding format profiles to a repository registry to inform preservation services. *IJDC*, 2(2):3–19, 2007.

[6] R. Graf and S. Gordea. Aggregating a knowledge base of file formats from linked open data. In *Proceedings of the 9th International Conference on Digital Preservation (iPres 2012)*, pages 293–294, Toronto, Canada, October 1-5 2012.

[7] S. Granger. Emulation as a digital preservation strategy. *D-Lib Magazine*, Vol. 6 (10), 2000. `http://www.dlib.org/dlib/october00/granger/10granger.html`.

[8] T. O. Group. *ArchiMate 2.0 Specification*. Van Haren Publishing, 2012.

[9] T. S. S. Institute. Approaches to software sustainability. Website. `http://www.software.ac.uk/resources/approaches-software-sustainability`.

[10] D. B. Marcum. The preservation of digital information. *The Journal of Academic Librarianship*, 22(6):451 – 454, 1996.

[11] R. Mayer and A. Rauber. Towards Time-resilient MIR Processes. In *Proceedings of the 13th International Society for Music Information Retrieval Conference (ISMIR 2012)*, pages 337–342, Porto, Portugal, October 8-12 2012.

[12] R. Mayer, A. Rauber, M. A. Neumann, J. Thomson, and G. Antunes. Preserving scientific processes from design to publication. In P. Zaphiris, G. Buchanan, E. Rasmussen, and F. Loizides, editors, *Proceedings of the 16th International Conference on Theory and Practice of Digital Libraries (TPDL 2012)*, volume 7489 of *Lecture Notes in Computer Science*, pages 113–124, Cyprus, September 23–29 2012. Springer.

[13] W. OWL Working Group. *OWL 2 Web Ontology Language: Document Overview*. W3C Recommendation, 27 October 2009. Available at `http://www.w3.org/TR/owl2-overview/`.

[14] J. Rothenberg. *Avoiding Technological Quicksand: Finding a Viable Technical Foundation for Digital Preservation*. Council on Library and Information Resources, January 1999.

[15] R. Treinen and S. Zacchiroli. Description of the CUDF Format. Technical report, 2008. http://arxiv.org/abs/0811.3621.

[16] J. van der Hoeven, B. Lohman, and R. Verdegem. Emulation for digital preservation in practice: The results. *IJDC*, Vol. 2 (2):123–132, 2007.

[17] R. van Diessen. Preservation requirements in a deposit system. Technical report, IBM/KB Long-Term Preservation Study Report Series #3, IBM Netherlands, Amsterdam, 2002.

[18] C. Webb. *Guidelines for the Preservation of Digital Heritage*. National Library of Australia, 2005.

[19] I. Witten and E. Frank. *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, San Francisco, 2nd edition, 2005.

Networked Instruction for Research Data Curation Education: The CRADLE Project

Helen Tibbo
School of Information & Library Science
University of North Carolina at Chapel Hill
201 Manning Hall, CB# 3360
Chapel Hill, NC 27599-3360
919-962-8063
tibbo@ils.unc.edu

Thu-Mai Christian
School of Information & Library Science
University of North Carolina at Chapel Hill
229C Davis Library, CB# 3355
Chapel Hill, NC 27599-3355
919-962-6293
thumai@email.unc.edu

ABSTRACT

In this paper, we describe a new initiative to develop a massive open online course (MOOC) for training library and information science students, library practitioners, and data producers in data curation. The Curating Research Assets and Data using Lifecycle Education (CRADLE) project exploits the affordances of MOOC technology to provide a networked learning environment that will encourage and foster the creation of research ecosystems in which CRADLE participants—library and information graduate students, library practitioners, and data producers—will have opportunities to collaborate with and learn from others engaged in data curation practice.

General Terms

communities, case studies and best practice, training and education

Keywords

data curation, data curation education, data management, massive open online course, MOOC

…data scientists [including] librarians [and] archivists… have the responsibility to design and implement education and outreach programs that make the benefits of data collections and digital information science available to the broadest possible range of researchers, educators, students, and the general public. – National Science Board, 2005

If data curation is viewed as a means to advance science … then libraries need to partner closely with investigators in the sciences and in other disciplines they serve. Because data vary so much by field, and by investigator, generic approaches to data collection are not feasible. – Christine Borgman, 2010

1. INTRODUCTION

While "standing on the shoulders of giants" and building on centuries of discoveries and painstaking research, much of 21st Century physical, medical, and social sciences are radically different from their predecessors that revolved around observation, experimentation, and more recently, small-scale computation. Today's "e-Science" (Hey & Hey, 2006) or "data-intensive science" (Gray & Szalay, 2007) or what Jim Gray of Microsoft Research termed in 2007 "fourth paradigm" science, (Bell, Hey, and Szalay, 2009; Gray, 2007; Hey, Tansley, and Tolle, 2009; Microsoft Research, 2006) is increasingly "carried out through distributed global collaborations enabled by the Internet" (UKNESC, 2012). This science features use and significantly, re-use, of very large data collections, very large scale computing resources, and high performance visualizations (Borgman, 2007; Borgman, 2012; Carlson & Anderson, 2007; SCARP Project, 2009). The stakes are high as e-Science promises discoveries and benefits not possible with more traditional methodologies. Social scientists are also facing the challenges of large-scale data. King observes that the "massive increases in the availability of informative social science data are making dramatic progress possible in analyzing, understanding, and addressing many major societal problems. Yet the same forces pose severe challenges to the scientific infrastructure supporting data sharing, data management, informatics, statistical methodology, and research ethics and policy, and these are collectively holding back progress" (King, 2011, p. 719). Humanists have also taken up the data-intensive approach, and the term "cyberscholarship" refers to scholarly research using high performance computing and digital libraries (American Council of Learned Societies, 2006; Arms, 2008).

Despite the apparent focus on technology, today's research environment is not just about high-capacity networks and large-scale digital data storage. It is not just about creating terabytes of new data or analyzing arrays of existing data in new ways. Effective and efficient data lifecycle management lies at the heart of today's research enterprise (DCC, "What"; Lord, Macdonald, Lyon, & Giaretta, 2004). For example, if data are not adequately or accurately described using metadata they will not be found in data stores, be interoperable, or understood for re-use. If sensitive data are not de-identified or kept securely, privacy and confidentiality will be breached. Data-intensive science presents a wide array of data management challenges for researchers, information and computer scientists, librarians, and data archivists as well as universities and public and private research laboratories that create and house data (ARL, 2007; Borgman, 2008; Choudhury, 2008; Garritano & Carlson, 2009; Gold, 2007a; Gold, 2007b; Hey & Hey, 2006; Jones, 2008). For truly productive science and scholarship that maximizes every research dollar and makes the investment in data creation re-usable, researchers must work in concert with data managers and digital curators (Abbot, 2008; DCC, 2010; Joint, 2007; Swan & Brown, 2008; National Academy of Sciences, 2009).

As e-Science takes root, producing unprecedented volumes of data in various and novel data formats, associated research data management challenges have also propagated. These challenges have invaded the purview of library and information science (LIS) professionals who are being called upon to tend to them. Many believe that data curation aligns with both the library mission to collect and provide access to scholarly materials and the librarian expertise that includes metadata, archival preservation, and bibliographic citation—all of which are applicable to data curation (Shaffer, 2013; Harris-Pierce & Liu, 2012; Latham & Poe, 2012). Others, however, argue that data curation necessarily restructures library practices because of the incongruences between the level and type of technical skill and professional judgment required for dealing with data and that required for other types of library materials (Gold, 2007a; Salo, 2010).

These incongruences, according to Gold (2007b), are resolved when libraries gain "fluency across library and scientific cultures" (Building capacity and understanding, para. 2). Consequently, LIS graduate schools have developed data curation education programs to teach such fluency. These programs not only teach data curation concepts such as digital preservation and metadata, but also they recognize that students benefit most from learning these concepts within the context of the research communities that produce data. The data curation specialization offered by the Graduate School of Library and Information Science at the University of Illinois at Urbana-Champaign requires foundational courses based on the concept of *purposeful curation* that emphasizes the cultural context, unique characteristics, and frameworks of data production, management, and sharing, while also placing emphasis on practical field experiences (Palmer, Weber, Munoz, & Renear, 2013). Likewise, the University of North Carolina at Chapel Hill's School of Information and Library Science offers a post-master's certificate in Data Curation that requires students to complete independent study projects that give students practical experience in a work environment (University of North Carolina, 2014). Carlson et al. (2011) maintain the need for data information literacy (DIL) programs that give students the ability to interpret and analyze data beyond simply managing them, with course content grounded in the cultures and practices of disciplinary domains.

Our assertion is that data curation education programs need to go a step further to address e-Science trends that have obliged the scientific community to (re-)define cultures and practices around data production, management, and sharing (Gray, 2009). The abundance of data production, decentralization of data sources, and interdisciplinary collaboration have necessitated the development of new technological approaches to data management and dissemination that enhance knowledge sharing in the data-intensive research landscape (Bell, Hey, & Szalay, 2009; Gray, 2009; Hey & Trefethen, 2003). If data curation education programs are to remain responsive to the rapidly evolving needs of the scientific community, programs need to adopt parallel approaches for the training and mobilization of LIS professionals who will be expected to apprehend the context in which research data are produced, managed, and disseminated. Therefore, data curation education must not only teach students the requisite data curation concepts defined in established graduate curricula, but also they must situate students within the relevant contextual framework.

2. THE CRADLE PROJECT

The IMLS-funded *Curating Research Assets and Data using Lifecycle Education* (CRADLE) project is working to take this step by developing a massive open online course (MOOC) that will provide instruction on data curation principles while focusing squarely on learning through networks of data management education and practice. A noteworthy outcome of e-Science has been the creation of "research ecosystems" that exploit advances in Internet communications technology (Goodman & Wong, 2009). These research ecosystems have given the citizen scientist opportunities to make important contributions to the corpus of scientific discovery, offered flexibility that has enabled interdisciplinary collaborations for solving large-scale problems, and provided access to tools that make scientific data comprehensible to a broader audience of individuals with varying levels of expertise (Goodman & Wong, 2009).

Likewise, the MOOC platform will allow CRADLE participants to exploit the same technological affordances to promote and support learning in a networked environment. Learners will be given access to the necessary technology and tools to enable them to construct similar research ecosystems in which individuals will be able to engage with and learn from others involved in data curation practice and make contributions to greater discussions around data curation. CRADLE will not only teach librarians the skills required for preparing data for long-term preservation and use, but also foster knowledge ecosystems by:

- Assigning projects that require students to make contact with data producers and information professionals at their local universities, libraries, research centers, or data repositories;
- Hosting virtual summits for CRADLE graduates that provide ongoing opportunities to share data management experiences and continue engagement with data management issues;
- Sponsoring opportunities for CRADLE students and graduates to attend data management symposia that feature significant players across the data management landscape; and
- Establishing virtual sandboxes and other technology that enable students to collaborate on data management challenges, with each student assigned to different data management stakeholder roles.

While individual CRADLE learning modules on data curation topics will contain content aimed toward specific audiences—LIS students, library practitioners, and data producers—each type of individual will interact with one another to solve data management problems. These interactions will encourage and foster an environment in which they can seed networks, which will grow as students also engage with their local research communities to explore first-hand the challenges of data management.

Moreover, CRADLE will provide an environment that will aid in the alignment of efforts to promote standards of data curation practice and to shift the culture toward one that recognizes research data as valued assets essential to the sustainability of the research enterprise. Where LIS students, library practitioners, and data producers coalesce on solutions to data management problems, discoveries of commonalities in data culture and practices may inform the establishment of best practices and encourage their adoption. CRADLE will serve as the backdrop from which effective data management education and best practice will emerge. The dynamic and unpredictable nature of research in the *fourth paradigm* (Gray, 2009) requires a more profound engagement with the research community to allow data curation education to adapt accordingly. No longer can information professionals operate within the confines of deep-seated archival principles and practices; librarians and archivists must find station within research ecosystems populated by data management stakeholders.

3. CONCLUSION

As current programs and novel initiatives such as CRADLE continue to develop and evolve, further study will be necessary to determine their success in preparing the next generation of librarians and information professionals as well as researchers themselves in meeting data management requirements of funders, journal publishers and institutions. If and when these data curation programs are proven successful, "working with data will become a mature component of librarianship when it is accepted into regular library practices; when terms like 'data reference' become simply 'reference' and datasets are not given any more specific or specialized treatment than other library collections" (Witt, 2012, p. 186). For this to happen, librarians must become active participants in the research community, making meaningful connections to individuals confronting data challenges, and arriving at common solutions for overcoming those challenges.

4. ACKNOWLEDGMENTS

The CRADLE project is sponsored by the Institute of Museum and Library Services, under award #RE-06-13-0052-13

5. REFERENCES

[1] American Council of Learned Societies Commission on Cyberinfrastructure for the Humanities and Social Sciences. (2006). *Our cultural commonwealth* (No. The report of the American Council of Learned Societies Commission on Cyberinfrastructure for the Humanities and Social Sciences). New York, NY: American Council of Learned Societies. Retrieved from http://www.acls.org/uploadedFiles/Publications/Programs/Our_Cultural_Commonwealth.pdf

[2] Arms, W. Y. (2008). Cyberscholarship: High performance computing meets digital libraries. *The Journal of Electronic Publishing*, *11*(1). doi:10.3998/3336451.0011.103

[3] ARL Joint Task Force on Library Support for E-Science. (2007). Agenda for developing e-Science in research libraries (Final Report and Recommendations). Washington, D.C.: Association of Research Libraries Retrieved from http://old.arl.org/bm~doc/ARL_EScience_final.pdf

[4] Atkins, D. (2003). *Report of the National Science Foundation Blue-Ribbon Advisory Panel on Cyberinfrastructure* (No. cise051203). Arlington, VA: National Science Foundation. Retrieved from http://www.nsf.gov/od/oci/reports/atkins.pdf

[5] Bell, G., Hey, T., & Szalay, A. (2009). Beyond the data deluge. *Science*, *323*(5919), 1297–1298. doi:10.1126/science.1170411

[6] Borgman, C. L. (2007). Scholarship in the digital age : information, infrastructure, and the Internet. Cambridge, Mass.: MIT Press.

[7] Borgman, C. L. (2008, June 27). *The role of librarians in e-science*. Conference Presentation presented at the European Conference of Medical and Health Libraries, Helsinki, Finland. Retrieved from http://blip.tv/eahil2008/the-role-of-libraries-in-e-science-christine-borgman-eahil2008-1045049

[8] Borgman, C. L. (2012). The conundrum of sharing research data. *Journal of the American Society for Information Science and Technology*, *63*(6), 1059–1078. doi:10.1002/asi.22634

[9] Carlson, S., & Anderson, B. (2007). What are data? The many kinds of data and their implications for data re-use. *Journal of Computer-Mediated Communication*, *12*(2). Retrieved from http://jcmc.indiana.edu/vol12/issue2/carlson.html

[10] Carlson, J., Fosmire, M., Miller, C. C., & Nelson, M. S. (2011). Determining data information literacy needs: A study of students and research faculty. *Portal: Libraries and the Academy*, *11*(2), 629–657.

[11] Choudhury, G. S. (2008). Case study in data curation at Johns Hopkins University. *Library Trends*, *57*(2), 211–220. doi:10.1353/lib.0.0028

[12] Digital Curation Centre (DCC). (n.d.). What is digital curation? Retrieved April 11, 2014, from http://www.dcc.ac.uk/digital-curation/what-digital-curation

[13] Digital Curation Centre (DCC). (2010). Resources for Digital Curators. *Introduction to Curation*. Retrieved from http://www.dcc.ac.uk/resources/briefing-papers/introduction-curation

[14] Garritano, J., & Carlson, J. (2009). A subject librarian's guide to collaborating on e-Science projects. *Issues in Science and Technology Librarianship*, *57*(Spring 2009). doi:10.5062/F42B8VZ3

[15] Gold, A. (2007a). Cyberinfrastructure, data, and libraries, part 1: A cyberinfrastructure primer for librarians. *D-Lib Magazine*, *13*(9/10). doi:10.1045/september20september-gold-pt1

[16] Gold, A. (2007b). Cyberinfrastructure, data, and libraries, part 2: Libraries and the data challenge: Roles and actions for libraries. *D-Lib Magazine*, *13*(9/10). doi:10.1045/september20september-gold-pt2

[17] Goodman, A., & Wong, C. (2009). Bringing the night sky closer: Discoveries in the data deluge. In T. Hey, S. Tansley, & K. Tolle (Eds.), *The fourth paradigm: Data-intensive scientific discovery* (pp. 39–44). Redmond, WA: Microsoft Research. Retrieved from http://research.microsoft.com/en-us/collaboration/fourthparadigm/4th_paradigm_book_part1_goodman_wong.pdf

[18] Gray, J. (2009). Jim Gray on e-Science. In A. J. G. Hey, S. Tansley, & K. M. Tolle (Eds.), *The fourth paradigm: Data-intensive scientific discovery*. Redmond, WA: Microsoft Research.

[19] Gray, J., & Szalay, A. (2007, January 11). *eScience--A transformed scientific method*. Presented at the Computer Science and Technology Board of the National Research Council, Mountain View, CA. Retrieved from http://research.microsoft.com/en-us/um/people/gray/talks/NRC-CSTB_eScience.ppt

[20] Harris-Pierce, R. L., & Liu, Y. Q. (2012). Is data curation education at library and information science schools in North America adequate? *New Library World*, *113*(11), 598–613. doi:10.1108/03074801211282957

[21] Hey, A. J. G., & Trefethen, A. (2003). The data deluge: An e-science perspective. In F. Berman, G. Fox, & A. J. G. Hey (Eds.), *Grid Computing: Making the Global Infrastructure a Reality* (pp. 809–824). New York: Wiley.

[22] Hey, T., & Hey, J. (2006). e-Science and its implications for the library community. *Library Hi Tech*, *24*(4), 515–528. doi:10.1108/07378830610715383

[23] Joint, N. (2007). Data preservation, the new science and the practitioner librarian. *Library Review*, *56*(6), 451–455. doi:10.1108/00242530710760337

[24] Jones, E. (2008). *e-Science talking points for ARL deals and directors*. Washington, D.C.: Association of Research Libraries. Retrieved from http://www.arl.org/storage/documents/publications/e-science-talking-points.pdf

[25] King, G. (2011). Ensuring the data-rich future of the social sciences. *Science*, *331*(6018), 719–721. doi:10.1126/science.1197872

[26] Latham, B., & Poe, J. W. (2012). The library as partner in university data curation: A case study in collaboration. *Journal of Web Librarianship*, *6*(4), 288–304. doi:10.1080/19322909.2012.729429

[27] Lord, P., Macdonald, A., Lyon, L., & Giaretta, D. (2004). From data deluge to data curation. In *Proceedings of the 3rd UK eScience All Hands Meeting* (pp. 371–375). Nottingham, UK: Citeseer. Retrieved from http://www.ukoln.ac.uk/ukoln/staff/e.j.lyon/150.pdf

[28] Microsoft Research. (2006). 2020 Science. Retrieved April 11, 2014, from http://research.microsoft.com/en-us/um/cambridge/projects/towards2020science/background_overview.htm

[29] National Academy of Sciences. (2009). Ensuring the integrity, accessibility, and stewardship of research data in the digital age. Washington, DC: The National Academies Press. Retrieved from http://www.nap.edu/catalog.php?record_id=12615#description

[30] Palmer, C. L., Weber, N. M., Munoz, T., & Renear, A. H. (2013). Foundations of data curation: The pedagogy and practice of "purposeful work" with research data. *Archive Journal*, (3). Retrieved from http://www.archivejournal.net/issue/3/archives-remixed/foundations-of-data-curation-the-pedagogy-and-practice-of-purposeful-work-with-research-data/

[31] Salo, D. (2010). Retooling libraries for the data challenge. Ariadne, 63. Retrieved from http://www.ariadne.ac.uk/issue64/salo

[32] SCARP Project. (2009). *Disciplinary approaches to sharing, curation, reuse and preservation* (Final Report). Bristol, UK: JISC. Retrieved from http://www.dcc.ac.uk/sites/default/files/documents/scarp/SCARP-FinalReport-Final-SENT.pdf

[33] Shaffer, C. (2013). The role of the library in the research enterprise. *Journal of eScience Librarianship*, *2*(1), 8–15. doi:10.7191/jeslib.2013.1043

[34] Swan, A., & Brown, S. (2008). The skills, role and career structure of data scientists and curators: An assessment of current practice and future needs. UK: JISC. Retrieved from http://ie-repository.jisc.ac.uk/245/1/DataSkillsReport.doc

[35] United Kingdom National e-Science Centre (UKNESC). (2012). Defining e-Science. Retrieved April 11, 2014, from http://www.nesc.ac.uk/nesc/define.html

[36] University of North Carolina (2014). Post master's certificate: Data Curation. *UNC School of Information and Library Science*. Retrieved from http://sils.unc.edu/programs/graduate/post-masters-certificates/data-curation

[37] Witt, M. (2012). Co-designing, co-developing, and co-implementing an institutional data repository service. *Journal of Library Administration*, *52*(2), 172–188. doi:10.1080/01930826.2012.655607

Addressing the personal digital archives needs of a contemporary artist

Sam Meister
University of Montana
32 Campus Dr.
Missoula, MT 59812
406-243-4036
sam.meister@mso.umt.edu

ABSTRACT

The study of personal digital archives is an emerging research area that crosses multiple domains including digital preservation, personal information management, archives, and recordkeeping. The practical need and desire for solutions and tools to meet the long-term preservation and access needs of individuals is increasing in relation to the ubiquitous production of digital information in contemporary society. To date, most digital preservation research has focused on the development of methods, tools, and solutions for institutional contexts such as libraries, archives, and other types of repositories. The personal context of an individual is distinct from organizational or institutional contexts, and necessitates new methods and approaches to better understand and develop solutions to meet personal digital recordkeeping and preservation needs. This paper describes a research project in progress that is focusing on this personal context by utilizing a case study approach to explore the design, development and implementation of personal digital recordkeeping system for a specific type of individual: a contemporary artist.

General Terms

Communities, specialist content types, case studies and best practice,

Keywords

Personal digital archives, recordkeeping, artist records

1. INTRODUCTION

The study of personal digital archives is an emerging research area that crosses multiple domains including digital preservation, personal information management, archives, and recordkeeping. The practical need and desire for solutions and tools to meet the long-term preservation and access needs of individuals is increasing in relation to the ubiquitous production of digital information in contemporary society. To date, most digital preservation research has focused on the development of methods, tools, and solutions for institutional contexts such as libraries, archives, and other types of repositories. The personal context of an individual is distinct from organizational or institutional contexts, and necessitates new methods and approaches to better understand and develop solutions to meet personal digital recordkeeping and preservation needs. This paper describes a research project in progress that is focusing on this personal context by utilizing a case study approach to explore the design, development and implementation of personal digital recordkeeping system for a specific type of individual: a contemporary artist.

2. PREVIOUS WORK

2.1 Personal digital archives

In a series of articles Marshall [1, 2] reports on studies conducted with a wide array of "consumers" to learn more about their understanding, behavior, and actions in relation to personal digital archives. From interviews and direct observations of personal digital archiving activities of a number of different types of individuals, Marshall describes a range of categories of personal digital archiving principles that consumers practice and the related challenges associated with those principles from the perspective of designing a personal digital archiving service. In distilling the findings of her studies, Marshall articulates the notion of "benign neglect" as being inherent to the practice of most individuals, and the need for the recognition of this phenomenon within the design of any future personal digital archiving service. While positioned within the personal information management field, Marshall's work crosses into the domain of archives, recordkeeping, and digital preservation, and represents an important contribution to an understanding of the personal digital archiving needs of the general consumer.

Lee and Capra [3] further explore the intersections between personal information management and archival literature. In their analysis of the commonalities between personal information management and archives and records management theories and practice, Lee and Capra recommend future research areas in which to further explore these connections, including "designing systems that are attentive to individual needs and behaviors" and "individual scale digital preservation". In a similar vein to Marshall's findings, these recommendations represent a shift from studies focused on identifying and understanding personal digital archiving behaviors and needs towards projects that explore the design and development of systems and services that address those needs.

Within the archival literature a number of studies and projects have investigated personal digital archives from the perspective of collecting institutions, the places where personal digital archives may eventually be acquired and managed. The Digital Lives project [4] is a recent significant contribution to this perspective, developing an "intellectual framework to help to better understand how people create, organize, manage, use and dispose of their personal digital archives" based on interviews with multiple stakeholders including creators and curators. Additional studies [5,6] have explored the personal digital archiving habits, behaviors, and actions of specific types of creators including writers and photographers. Findings from these studies include recommendations that can be characterized as a need for increased interactions between archivists and creators before digital materials are acquired or transferred to an archive. These recommendations include archivists providing guidance or simple steps for creators to follow in creating and managing digital archives, actions which will potentially benefit an archivist or institution that may acquire the materials in the future.

Cunningham's [7] position on personal digital archives, or in his words, "personal recordkeeping", functions as an early statement in support of archivists providing guidance to individual creators, but also suggests a more fundamental shift for the role of the archivist. In alignment with the records continuum conceptual model, Cunningham suggests that archivists should endeavor to be more directly involved in the records creation process to ensure that records are, "created and captured into well-designed, well-documented recordkeeping systems". He further articulates that recordkeeping standards and methodologies, such as ISO 15489 [8], based on the records continuum conceptual framework, could potentially be applicable in the realm of personal recordkeeping and personal digital archives, and should be tested through "research projects with some individual creators". Cunningham's embrace of the records continuum approach to personal digital archives represents a distinct contrast to perspectives based on the life cycle model which suggest increased, but still fairly limited, intervention on the part of the archivist in relation to the records creation process.

2.2 Artist records

Within the larger context of the InterPARES 2 project, researchers investigated the documentation practices of the performance artist Stelarc [9]. Utilizing a case study approach, through interviews with the artist and observations of performances, researchers analyzed the creative activities of the artist to determine how the digital entities that resulted from these activities corresponded to a traditional definition of a record based on diplomatics theory. With recognition of its limitations [10], the Stelarc case study represents an important contribution to understanding the records creation process of a contemporary artist and the related recordkeeping and preservation system requirements.

Another contribution to the understanding of the recordkeeping and archival needs of contemporary artists is the recent project, Studio Archives: Voices of Living Artists, Their Assistants, and Their Archivists [11]. Through a series of interviews with a range of artists at various career stages, this project has attempted to document the current state of artist's studio archives, articulate recordkeeping challenges and needs, and build relationships between artists and information professionals. The primary goal of this project is to produce, "a guide for artists on how to establish an archive, and how to maintain it over time". To date, this guide has not been published, but is intended to include guidance for managing and preserving digital content.

Furness [12] has produced one of the few examples of a study that investigated the recordkeeping practices of a single contemporary artist using an exploratory case study approach. Specifically, this project sought to, "understand, through empirical investigation, the many factors that shape the artist's recordkeeping and archives in the personal sphere and contribute to the nature of the eventual archival fonds in the institution". In this case, the artist records had already been acquired by an archival institution. Through interviews with the artist, Furness investigated the relationship between the artist's recordkeeping activities and her creative practice. Additionally, the archivist responsible for the acquisition of the artist's records was interviewed to understand the archival transfer process. While this project is framed by the context of an institutional archive, it offers valuable insight into the creative process of a contemporary artist and the resulting records of that process.

3. RESEARCH QUESTIONS

The current project described here seeks to build on contributions to the understanding of the issues, challenges, and opportunities related to the emerging field of personal digital archives provided in the studies described in the previous section. This project seeks to move beyond current understanding, to extend and expand the discourse on personal digital archives by engaging in a project with a very practical goal: the design, development, and implementation of a functional personal digital recordkeeping system for a contemporary artist. While the primary goal of the project is to produce a recordkeeping system that integrates digital preservation actions throughout the artist's creative process, a secondary goal is to support the preservation of the outputs of this process, including artworks that are or include digital objects as elements. The research questions that propel this project are as follows:

1. What are the personal recordkeeping and digital preservation needs of a contemporary artist?

2. What are the specific recordkeeping and preservation system requirements?

3. Can these requirements be met by current recordkeeping and digital preservation software and services?

4. What does a functional personal recordkeeping system for a contemporary artist look like?

4. METHODS AND APPROACH

To investigate these research questions, the project is utilizing a case study approach to explore the development of a recordkeeping system for an individual contemporary artist. The case study approach provides a framework to focus research on a case unit to allow for a more intensive and detailed investigation than may be possible with multiple units [13]. The current project's focus on a single contemporary artist as the case unit is intentional and corresponds to a hypothesis that artists are a specific type of creator that will have specific recordkeeping needs related to their creative process. The case study approach provides a framework for a rich and deep exploration of the artists' creative process, and the development and implementation of a recordkeeping system upon which to assess this initial hypothesis.

Within a case study structure the project is also employing an action research method to facilitate the process of working towards very practical goals and objectives. Action research provides a structure for projects that entail a different approach from the traditional model of researcher as an observer of research subjects. In action research, the researcher role functions more as a facilitator and the traditional role of the subject is instead an active participant in all phases of the research project that seeks to investigate, plan, and implement actual change [14]. The action research method offers a structure that is particularly applicable to personal digital archives research, in that it entails an iterative process of planning, acting, and evaluating to work towards project goals. Previous research [15] has illustrated the idiosyncratic nature of how individuals think, behave, and act in relation to personal digital recordkeeping. In the context of the current project, which is based on collaboration between an archivist and an artist, an iterative approach that incorporates continual evaluation will allow for flexibility in adjusting project activities in relation to potential nuanced discoveries, as well as assist in ensuring that both participants are making progress towards stated goals.

Finally, the current project is also investigating the applicability of the records continuum conceptual framework to the realm of personal digital recordkeeping. Situated outside of a traditional institutional context of an archivist acquiring the records of a donor for inclusion in an institutional collection, the current project aspires to explore the potential role of an archivist in engaging and collaborating with creators to develop practical solutions for their personal recordkeeping, digital preservation and access needs.

5. PROJECT STATUS

The project is currently in an early phase. To date, work completed includes establishing a relationship with the specific contemporary artist, and early steps in describing and mapping the creative process of the artist.

5.1 Establishing the relationship

The decision to design the project around the recordkeeping needs of a contemporary artist did not follow a traditional process of the researcher selecting a particular research subject based on the subject's qualities in relation to a specific set of research questions. Instead, the relationship was established and evolved through a series of social interactions in which the topic of personal digital archives was repeatedly discussed between the author and the contemporary artist. It is important to note that these interactions took place before the author was affiliated with their current institution, and evolved into a consultant (author) and client (artist) relationship as initial project goals and objectives were developed. This relationship has evolved into an equal collaboration as the project has shifted into a more formal mode, including the development of the specific research questions and the decision to utilize the research methods articulated in the previous section.

5.2 Creative process mapping

Through a series of semi-structured interviews the author and artist are engaging in the process of developing a set of diagrams that visually map the various steps and activities that are involved in creation of a typical artwork or project. The creative process diagrams will be revised multiple times through an iterative process of collaborative review between the author and the artist. The final versions of the creative process diagrams will include the identification of specific digital objects that function as outputs of the various creative process activities.

6. FUTURE WORK

Additional proposed project phases include:

1. Identification of specific digital objects as records and determination of record value

2. Design and development of recordkeeping system requirements

3. Identification and testing of tools to meet system requirements

4. Implementation initial version of recordkeeping system

5. Assessment of recordkeeping system functionality and use

6. Modification of recordkeeping system elements based on assessment results

7. REFERENCES

[1] Marshall, Catherine C. 2008. Rethinking Personal Digital Archiving, Part 1: Four Challenges from the Field. *D-Lib Magazine (14)* 3/4 http://www.dlib.org/dlib/march08/marshall/03marshall-pt1.html.

[2] Marshall, Catherine C. 2008.Rethinking Personal Digital Archiving, Part 2: Implications for Services, Applications, and Institutions. *D-Lib Magazine (14)* 3/4 http://www.dlib.org/dlib/march08/marshall/03marshall-pt2.html.

[3] Lee, C. A. & Capra, R. 2011. And now the twain shall meet: Exploring the connections between PIM and Archives. In C.A. Lee, *I, Digital: Personal collections in the digital era.* (pp. 29 – 77) Chicago: Society of American Archivists.

[4] Williams, P., John, J. L., & Rowland, I. 2009. The personal curation of digital objects: A lifecycle approach. In *Aslib Proceedings* (Vol. 61, No. 4, pp. 340-363). Emerald Group Publishing Limited.

[5] Becker, D., & Nogues, C. 2012. Saving-Over, Over-Saving, and the Future Mess of Writers' Digital Archives: A Survey Report on the Personal Digital Archiving Practices of Emerging Writers. *American Archivist*, *75*(2), 482-513.

[6] Spurgin, K. M. 2011. "Three backups is a minimum": A first look at norms and practices in the digital photo collections of serious photographers. In C.A. Lee, *I, Digital: Personal collections in the digital era.* (pp. 151 – 201) Chicago: Society of American Archivists.

[7] Cunningham, A. 1999. Waiting for the ghost train: strategies for managing personal electronic records before its too late, *Archival Issues*, *(24)*1, pp. 55-64.

[8] BSI, *BS ISO 15489-1: Information and documentation – Records Management – Part 1: General*. BSI, London; BSI (2001), *PD ISO/TR 15489-2: Information and Documentation – Records Management – Part 2: Guidelines*, BSI, London, 2001.

[9] Daniel, Henry, and Cara Payne. 2007.*Case Study 02 Final Report: Performance Artist Stelarc*. InterPARES 2 Project: International Research on Permanent Authentic Records in Electronic Systems. Vancouver, BC: University of British Columbia. http://www.interpares.org/display_file.cfm?doc=ip2_cs02_final_report.pdf.

[10] Lau, A. J. 2013. Collecting Experiences. (Doctoral dissertation, University of California, Los Angeles, 2013) http://www.escholarship.org/uc/item/9f8572zp

[11] Gendron, H. & Imm-Stroukoff, E. 2011. Studio Archives: Voices of living artists, their assistants, and their archivists. Paper presented at *Artist records in the archives symposium.* New York, NY.

[12] Furness, A. L. 2012. Towards a Definition of Visual Artists' Archives:Vera Frenkel's Archives as a Case Study (Doctoral dissertation, University of Toronto, 2012). http://hdl.handle.net/1807/32714

[13] Denzin, N. K., & Lincoln, Y. S. (Eds.). 2011. *The SAGE handbook of qualitative research.* Sage.

[14] Stringer, E. T. 2013. *Action research.* Sage.

[15] Marshall, C. C., Bly, S., & Brun-Cottan, F. 2006. The long term fate of our digital belongings: Toward a service model for personal archives. In *Archiving Conference* (Vol. 2006, No. 1, pp. 25-30). Society for Imaging Science and Technology

Virtualisation as a Tool for the Conservation of Software-Based Artworks

Patrícia Falcão
Time Based Media Conservation, Tate
7-14 Mandela Way
London SE1 5SR
+44 20 7887 8574
patricia.falcao@tate.org.uk

Alistair Ashe
Time Based Media Conservation, Tate
7-14 Mandela Way
London SE1 5SR
+44 7891 296 532
alistair.ashe@tate.org.uk

Brian Jones
Information Systems, Tate
20 John Islip Street
London SW1P 4RG
+44 020 7887 8505
brian.jones@tate.org.uk

ABSTRACT

Tate has a small but growing collection of software-based artworks. From the outset basic preservation procedures, like testing equipment, backing up hard-drives and assets or thoroughly documenting the hardware and software were put in place, but it was clear that these procedures would need revising over time and as our experience grew. Tate's earliest software-based artwork was created in 2003 and after 10 years the issues around aging technologies are becoming more obvious and new strategies for preservation are more urgently needed. The number of artworks being acquired and displayed is increasing and therefore better workflows must be developed to accommodate this increase.

This paper describes a short project to scope the use of virtualisation for preserving software-based artworks in Tate's Collection. It briefly explains the tests performed, in terms of the techniques, resources and expertise involved.

Through the tests it was confirmed that virtualisation is a viable strategy for the preservation of software-based artworks, and that it meets our requirement that the artworks be stored as a complete system independent from the original hardware. It was also a main requirement that different virtualisation tools must support the resulting virtual machines. As a conclusion, the workflow currently being developed for the preservation of Tate's software-based artworks will be outlined.

General Terms

Case studies and best practice, communities

Categories and Subject Descriptors

H.3 [**Information Storage and Retrieval**]: Digital Libraries
J.5 [**Arts and Humanities**]: Arts, fine and performing

Keywords

Software-based Art, computer-based art, born-digital artworks, virtualisation, digital preservation, long-term access, museums.

1. INTRODUCTION

This paper discusses a current proposal for virtualising the software-based artworks in Tate's Collection as a preservation strategy. The proposal is defined within the context of the Tate Collection; the current strategies in place for the preservation of software-based artworks, the existing infrastructure and the resources available.

The proposal is the result of a short project[1] in which different departments at Tate worked collaboratively, namely: Collection Care Research, Conservation and Information Systems (IS). This research was done in the context of the Pericles project [1]. The team of researchers set out to create virtual machines (VMs) for two artworks as an experiment to test whether virtualisation is a viable strategy for preservation. The main aim of the experiment was to understand if these artworks could be virtualised and if, once virtualised, the software behaviour remained unchanged.

Based on prior work done on identifying risks for the preservation of these works [2], the team knew features such as processor speed, interfaces with peripherals or connection to the Internet could become problematic, and so the artworks tested included these functions.

Virtualisation has not been previously used for preservation at Tate, however, because Tate's IS infrastructure is based on virtualised servers, there is already the required expertise in-house. The project was an opportunity to start thinking about the preservation of software within the context of Tate's infrastructure, the current preservation practices and most importantly in cooperation with our IS department.

Because VMs are susceptible to obsolescence, creating one to run the artworks must be seen as an initial step. The virtualisation process copies the original operating system and artwork software to a form independent from the existing aging hardware. The existence of the complete system as a file that can be run in the current virtualisation platform, and benefits from the maintenance provided by the IS department, is a promising start. Also positive is the fact that the files created can be saved in a standard format, increasing the likelihood that the virtual machines remain sustainable for a longer period of time.

For the research project, the team tried two virtualisation tools, VMware [3] and VirtualBox [4]. We were interested in questions of long-term sustainability, whether virtualisation could create a functioning copy of the artworks and also how feasible it would be for a non-IT specialist to use the virtualisation tools. Once these issues were investigated, the next question was whether virtualisation could be made part of the preservation workflows currently in place.

The result of the research was the agreement that virtualisation is a valid strategy and that the tools tested partially match the following requirements:

[1] 20 days were allocated across staff in three different departments.

a) The artwork's software is able to run in a form that is as close as possible to the original system.

b) The virtualisation tools support current operating systems, like Windows XP.

c) The virtualisation tool is able to connect easily to different peripherals and networks.

d) The virtualisation tool is easy to use by non-programmers.

e) The resulting files are in a standard format supported by different virtualisation tools.

In preparation for the project, we tried to identify possible emulators to be tested alongside the virtualisation tools. We found some emulators for x86 platforms, such as Dioscuri [5] or QEMU [6]. Testing QEMU would have been interesting, but given the limited amount of time available we opted to test tools that we were already familiar with. Testing QEMU would be the next step in the research.

Overall the outcome of the entire project was positive. Virtualisation does make the artwork independent from the original physical machine, therefore reducing the risks related to hardware failure. It became clear that the virtualisation process is straightforward for some works, but that when peripherals are involved issues will arise that will need specialised intervention.

What we suggest is that by keeping software-based artworks running in virtual machines they are rendered less dependent on the original hardware. Once the original hardware stops working it should still be possible to revert to the virtual version. By identifying and documenting the significant properties of the artwork and the artwork software, and by ensuring that the virtual version is an exact copy of the original, we can then use the virtual version of the artwork to compare any new version that may be created. The new version could either be created on new hardware or migrated to a new operating system or programming language.

2. LITERATURE REVIEW

The importance of preserving the functionality of digital files and software was identified by Jeff Rothenberg in his 1995 paper *Ensuring the Longevity of Digital Documents*, within which he proposes using hardware emulation for the preservation of software environments where digital objects were originally created. Around 2000, the CEDARS project highlighted the importance of emulation as a preservation strategy. As a consequence, Leeds University, one of the main partners in CEDARS [7], developed DomesEm, an emulator for the BBC Domesday Project as part of another project, CAMILEON.[2]

Emulation applied explicitly to software-based artworks was first suggested by Richard Rinehart in 2002 [8][9], and in 2006 Rothenberg [10] published his report on the emulation of the interactive work *Erl King* by Roberta Friedman and Grahame Weinbren, where he describes the successful process of emulating the artwork, and illustrates the issues that arise around the process.

Since 2006, the emulation strategy has been further developed by the digital preservation community as a way of providing access to digital files. The most visible result is the DIOSCURI emulator, a collaboration between the National Library of the Netherlands [11], the PLANETS and KEEP projects [12] [13] and the company Tessella [14]. More recently QEMU has been adopted by the KEEP project in their Emulation Framework, but the tool itself was not developed specifically for preservation.

The proposal of 'emulation as a service' that is being developed by the project Baden-Wuerttemberg Functional Long-Term Archiving and Access (bwFLA) [15] sounds very promising. In the case studies published on their website there are examples of emulating artworks on CD-Rom. It would be interesting to test this emulation tool with different types of peripherals, as often required by software-based art installations.

The ZKM Case Studies [16], conducted as part of the project Digital Art Conservation, illustrate clearly the problems faced and resources needed when preserving and migrating artworks that are highly dependent on particular hardware.

Virtualisation has received less attention than emulation from the Digital Preservation community as a whole, but it has been suggested as a possible tool within the field of art conservation. One of the first references was in 2008, by Tabea Lurk and Juergen Enge [17], who tested the use of virtualisation for the conservation of software-based artworks. They defined the concept of *work logic* of an artwork as "the work logic identifies the core components of the artwork and describes the interlocking of the digital modules involved". They also describe the idea of *work relevant components* and *environmental elements*, and suggest creating an encapsulation layer around the *work relevant components* to maintain them. Enge gave the following informal definition of encapsulation in an interview with the PACKED project: "Encapsulation can mean just about anything that can provide a software layer between the artwork and, for instance the runtime environment, the operating system or the hardware." [18]

Emanuel Lorrain, in a case study for Mondophrenetic™ (2000, Herman Asselberghs, Els Opsomer, Rony Vissers), also suggests the use of virtualisation for the preservation of artworks. Lorrain points out several limitations found when using emulators: limited support for peripherals and more recent operating systems, dependence on voluntary work from a community, and therefore lack of reliability in the mid-term. He suggests using virtualisation, but again points to limitations for older operating systems. [19]

The key point that must be addressed to determine the successful outcome of an artwork's virtualisation is the comparison of the significant properties of the original and virtualised versions. The significant properties of commercial software were defined in the report *The Significant Properties of Software: A Study* [20]. Specific research on the significant properties of networked artworks was conducted at Tate in 2010. Kelli Dipple, Frederico Fazenda-Rodrigues and Pip Laurenson analysed the significant properties of networked artworks as part of the New Media Art Network on Authenticity and Performativity.[21] When comparing the categories identified in both reports it was easy to establish parallels. What also became clear was that software-based artworks require more granularity in describing user interaction, provenance and ownership and also functionality.

3. CONTEXT

[2] In February 2014 the webpages for both CEDARS and CAMILEON were no longer online.

Year Acquired	Year Produced	Artwork	Artist	Type	Operating System	Programming Language
2003	2003	Becoming	Michael Craig-Martin	Software, colour, monitor	Windows XP	2003 Director version (Executable) 2010 – Flash version (Executable)
2007	2005	Subtitled Public	Rafael Lozano-Hemmer	Software, interactive, colour, computer and video projections	Windows XP	Delphi (Executable)
2008	2007	Things Change	Michael Craig-Martin	3 wall-mounted LCD monitors and software	Windows XP	2003 Director version (Executable) 2010 – Flash version (Executable)
2009	2005	Limac shop	Sandra Gamarra	Installation with access to Website	Web based	Wordpress website (php, html, css, mysql database).
2010	2007	Brutalism: Stereo Reality Environment 3	José Carlos Martinat	Fibreboard, 3 printers, paper, tracking system, central processing unit, cables and web search program	Linux Ubuntu	Java (with mysql database)
2012	2005	Colors	Cory Arcangel	Software, Video, projection, colour and sound (stereo)	Mac OS	Objective C / C++
2012	2006	Astrophotography...The Traditional Measure of Photographic Speed in Astronomy...' by Siegfried Marx (1987)	Cerith Wyn Evans	Glass chandelier, flat screen and morse code unit	Windows XP	Commercial Software Morse Translator V12
2013	1997-2002	Adji- part of Library from Museum of Contemporary African Art	Meshac Gaba	Online Game	Web-based	Shockwave, Html, Css

Table 1. Software-based artworks in the Tate Collection

Artists have been using computers to produce artworks ever since software started being developed in the 1960s [22]. Since then, as in all parts of our culture, software has become another tool, amongst many, that artists use to create artworks.

Software-based art has its own circuit of festivals, like the Transmediale in Berlin [23] and collecting institutions, like Ars Electronica in Linz (which started as a festival but became a Centre in 1995)[24] and the ZKM- Centre for Art and Media Karlsruhe [25]. These are still very important hubs for the field. In the 1990s the mainstream contemporary art world also started to collect these types of works and nowadays, major art galleries, like the Lisson Gallery, London, sell software-based artworks by Cory Arcangel alongside sculptures by Ai Wei Wei, for example. Contemporary Art Museums also slowly began acquiring software-based artworks in the 90s and early 2000s.

In 2003, Tate acquired its first software-based artwork, *Becoming* (T11812*)* by Michael Craig-Martin. Since then another five works have been brought into the collection, and two more are in the process of being acquired. Table 1 lists these works.

Eight may seem like a small number of artworks to preserve, particularly if you compare it with around 400 Time-Based Media (TiBM) artworks, or the 70,000 works in the whole of the Tate Collection. Yet currently, the amount of resources needed to preserve one software-based artwork is much higher than the resource needed for an average TiBM artwork. This is due to the technical complexity of the artworks, their uniqueness, but also because the workflows required are not fully established.

The conservation section responsible for the preservation of software-based art at Tate is TiBM Conservation. This team, which is currently made up of eight members of specialist staff, is part of the Conservation Department. It is responsible for the conservation and installation of artworks using video, audio, film, slides, light-boxes, software and performance. TiBM Conservation was established as an independent section within the Conservation Department in 2004. The first TiBM conservator—

Pip Laurenson—was appointed within sculpture conservation in 1996.

The TiBM Conservation team is experienced at managing a variety of technologies and issues of obsolescence, and working with artists to understand both the requirements of an artwork, and the artist's attitude to change. For the TiBM Conservation team, software-based art is a challenge which is best met by both drawing on their existing experience of working with technology-based artworks; and by developing a new set of practices, technical skills and tools - and crucially an additional network of specialists.

Given the permanent technological evolution, and consequent broadening of the range of media in the collection, ongoing engagement in research and active collaboration is required to keep abreast of all the developments but also to devise new strategies to manage dying technologies.

Until recently, there was little collaboration between the conservation and the IS teams regarding the conservation of artworks, mostly because the conservation department had its own infrastructure and processes. This has changed significantly with the need to preserve high value digital information at all levels in the institution, not only for software-based artworks but also video and photography. We now have the opportunity to work with the IS department, who have made the engagement with conservation part of their strategic plan.

Institutions with different departmental structures, capacities and levels of institutional support have found other ways to ensure that they have the technical support they need. For example, some institutions have identified individuals who work on a freelance basis to act as the conduit between conservation and IS.

3.1. The Artworks

Software-based artworks are usually supplied to Tate on a computer, ready to be installed in the galleries.

Each of the eight software-based artworks has a programmed element that is bespoke; however, in the majority of cases the artist did not programme them themselves. The exception is *Colors* (2005), by Cory Arcangel an artist who does his own programming. The other artists in the collection have worked with a programmer to develop a system that will perform a series of actions. Some of the actions performed include: analysing video, mapping the location of visitors in a gallery, or displaying randomly composed tableaux of vector images.

All the artworks have different hardware requirements, and both computers and peripherals are usually supplied by the artist when the work is acquired.

The details of the Operating Systems, software elements and programming languages are provided in table 1. The table illustrates the variety of systems used, and why a network of specialists in different programming platforms is required.

4. CURRENT PRACTICE

4.1. Conservation Strategies

The goal of conservation is to ensure that artworks remain exhibitable within the defining parameters of the work, which are often tightly specified by the artists. To achieve this end, when an artwork comes into the collection, conservators work closely with the artist to identify the significant properties of the work and define what measures are appropriate for preservation. This is done by examining the artworks as they are supplied by the artist or gallery, requesting any additional information needed from the artist, and discussing (with the artist and programmer) what the issues for preservation are likely to be. At this stage possible preservation strategies are also discussed. The documentation of this process forms part of the artwork's conservation record.

Within the conservation of software-based artworks, to date, we have used three possible strategies:

1) Managed Storage- By keeping the hardware in good storage conditions and creating exhibition copies we are prolonging the life of the artwork in its original form.

2) Re-coding or replacing software elements – Conservators work closely with the artist/artist's programmer to re-code the work to a new platform. Another type of migration would be to replace one commercial software by another that has the same function, like a morse code translator in the work *"Astrophotography...The Traditional Measure of Photographic Speed in Astronomy...' by Siegfried Marx (1987)"* (2006) by Cerith Wyn Evans.

3) Virtualisation/Emulation- by this term we are referring to the creation of a virtual machine to run the original software, either by means of an emulation or virtualisation tool. In the next sections the virtualisation process is discussed in more detail.

The applicable strategies are dependent on the value attributed by the artist to a particular component of the artwork. For example, if a computer is designed by the artist and the object itself is the artwork, conceived of as a sculptural object, then storage is the only option. This is the case with Richard Hamilton's Diab DS-101 Computer (1985-9, T07124).

Some artists will define the source code as the artwork, for example Hans Diebner's Liquid Perceptron (2000), as documented by Tabea Lurk [17]. In this case the computer itself may be replaceable, but migration of the code is not an option. A combination of storage and virtualisation/emulation becomes the logical choice.

Across the artworks currently in the Tate collection, the software itself is predominantly considered a tool to produce a particular effect by the artist. Consequently, the preservation of this behaviour is identified as paramount rather than maintaining the original code. It is therefore appropriate to consider migration as a conservation strategy.

4.2. Significant Properties

When acquiring an artwork it is essential to identify its significant properties, as only by defining those is it possible to determine the best combination of conservation strategies to apply. In the *Final Report for the New Media Art Network on Authenticity and Performativity* [21] Kelly Dipple et al. categorises the possible Significant Properties of networked art as:

- Content and Assets
- Appearance
- Context
- Versions
- Formal and Structural Elements
- Behaviour
- Time
- Spatial or Environmental Parameters
- External Links or dependencies
- Function

- Processes
- Artist's Documentation of Process,
- Rules of engagement
- Visitor Experience
- Legal Frameworks

These significant properties are discussed in detail by P. Laurenson in *Old Media, New Media? Significant Difference and the Conservation of Software Based Art.* [26]

The significant properties will vary with the artwork, and are very closely related to the artist's intent. The same property can be significant or not depending on the value an artist attributes to that particular property. Defining them and finding ways of evaluating these properties in both the physical and virtual machines is in our opinion, the main challenge and the most important step in the process.

For *Becoming,* which is described in more detail in the next section, the artist's programmer, Daniel Jackson from AVCO, wrote a script to measure the speed at which images appear and disappear from the screen. For this one work this tool provides a concrete way to measure speed – a significant property of this work. However, this tool is specific to this artwork, and it is unlikely that its usefulness would be applicable to a different work. This indicates that there may be the need to write specific tools for other artworks as well. How to measure quantifiable significant properties is one further strand of research that we need to develop.

Further to the artist intent, it is also important to consider the technical history of the artwork. Migrating the software may be an option if we are presented with the loss of functionality of an artwork, but conservators must also consider the preservation of the production history of the artwork. Emulation or virtualisation may mean that the original program can be kept along with the functionality, and this would be a great advantage.

4.3. Existing Workflow

Before an artwork enters the Tate's collection Time-based Media Conservation creates a report containing a basic technical description of the artwork and a preservation plan. For this an initial discussion takes place with the different stakeholders; the artist and his programmer, curators, conservators and sometimes also technical staff in galleries.

Once the acquisition is approved and the actual work is received a more detailed analysis of the different software and hardware components is created.

At this point it is also standard practice to create an exhibition copy (on hardware as close to the original as possible) in house. This exhibition copy is created to protect the original from the wear and tear problems that arise when equipment is running for 70 hours a week in the museum gallery. In addition to these practical concerns, creating an exhibition format is also a way of understanding the work better, as issues always arise during the process of replication. This step also often requires the involvement of the artist/programmer, and is a moment when the initial description of the software is verified. By creating the exhibition copy we are reducing the risk of failure by wear and tear for the original systems of hardware and software.

In summary, the current approach for preservation focuses on two points:

Documenting the system and the artist's intent:

a) Creating and keeping system reports on the hardware and operating system, along with their specifications and any particular settings.

b) A narrative account of what the software does, and how it does it.

c) A stored copy of the source code, when the program is bespoke.

d) Communication with the artist and programmer about the artist's intent, significant properties, technological choices, preservation risks and the artist's preferences in terms of preservation. This process usually starts with an interview but then develops over time as needed.

Preserving the hardware:

a) Storing original hardware in appropriate environmental conditions.

b) Maintaining the equipment as required

c) Backing up hard-drives.

d) Creating an exhibition copy

Yet, these processes still leave the artworks highly dependent on particular equipment and therefore under threat of equipment failure. By virtualising the artwork, we can remove the dependency on physical hardware to reduce the risk of loss by equipment failure.

5. THE PROJECT

As previously stated, this collaborative twenty day research project was instigated to investigate virtualisation as a preservation tool for Tate's software based artworks.

Given the time-bound nature of the project, we opted to virtualise two artworks at each end of the complexity spectrum: *Becoming* (T11812), by Michael Craig-Martin; and *Brutalismo: Stereo Reality Environment 3* (T13251), by Jose Carlos Martinat Mendoza.

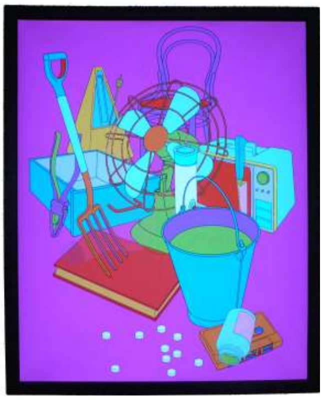


Figure 1-Michael Craig-Martin, *Becoming*, 2003

Becoming is a Windows XP executable that presents eighteen vividly coloured vector line drawings of everyday objects fading randomly, slowly in and out against a fuchsia pink background. It is presented in a custom-made monitor with an in-built computer.

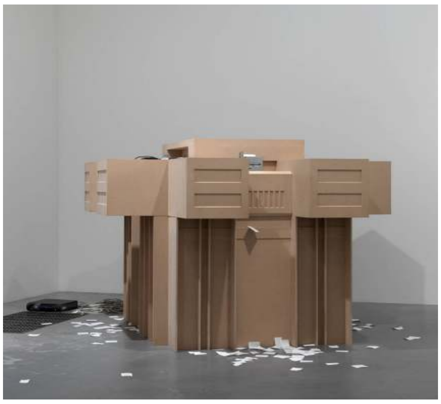


Figure 2- Jose Carlos Martinat Mendoza, *Brutalismo: Stereo Reality Environment 3,* 2007

Brutalismo is composed of both a sculptural and a software-based component. The work is described within Tate's online catalogue as follows: *"This sculpture is a scale model of the Peruvian military headquarters, an example of 'brutalist' architecture it was nicknamed the 'Pentagonito' (or 'little Pentagon'). During the Fujimori presidency, the building became notorious for the torture, murders and disappearances conducted by the secret service. The sculpture incorporates a computer which has been programmed to search the internet for references to 'Brutalismo / Brutalism', picking up extracts about Latin American and global dictatorships but also on architecture, forging associations between different kinds of 'brutalism' which it spews out onto the gallery floor."*[27]

Technically the work is composed of different software elements embedded in the shell of an Ubuntu operating system. It requires an internet connection to connect to Google and outputs to either RS232 or USB printers. We knew that interfaces with external systems cause the most problems for emulation and virtualisation, and therefore expected *Becoming* to be simple to virtualise, unlike *Brutalismo.*

Brian Jones, from Tate's IS department, created the test virtual machines using Virtual Box and VMware.

In considering the different tools the following criteria were used for evaluation :

- Sustainability- is the tool and industry standard or in widespread use?
- Interoperability-what file formats are supported, and are they supported by different platforms?
- Expertise and Infrastructure- is there any expertise in-house, and if so is it possible to use the existing infrastructure? Both tools support OVF, a widely adopted open standard for virtual machines.
- Costs- what are the costs involved?
- Features- can the tool create, open and convert virtual machines?
- Supported Operating Systems- which operating systems are supported?
- For how long has the tool been in use and development, is it a mature technology?

We chose Virtual Box because it is in widespread use and is a free, open source option, and VMware because it is the tool already being used by our IS department for the virtualisation of Tate's servers. They both allow the export of virtual machines in the Open Virtualisation Format (OVF), *a packaging standard designed to address the portability and deployment of virtual appliances* [28]. A further advantage of the OVF format is the metadata included, which describes the virtual machine's properties.

We started by creating a virtual machine, installing the operating system, and copying the software executable into the virtual machine. At Brian Jones' suggestion we then tried using a physical to virtual process, which proved to be more advantageous, as it captures all the information of the original system supplied by the artist, which often contains more than just the artwork.

A good example of where valuable information was retained by using the physical to virtual process was in the programming tools contained in one of the computers, which still contained the source code for the work, but also other code that had been adapted from other artworks. These are very interesting traces of how the software was developed. By looking at the programming tool we learned that the same source code had been used in a series of artworks, with minor adjustments. None of this information would have been captured if we had simply re-installed the software.

At the planning stage we expected problems when setting up the printers for *Brutalismo,* but because they already use a fairly recent type of connection, namely, a USB connection, this proved to be straightforward. It was also straightforward to open the OVF files created in VMWare using Virtual Box, which we had suspected could cause problems.

In discussions over the longevity of a virtual machine, we tried to identify the most likely cause of the virtual machine running Windows XP failing to run. We expect that VMware or any virtualisation platform will eventually stop supporting Windows XP, or 32-bit software. From these discussions it became clear that it is crucial to monitor the evolution of VMWare and the OVF format and their continued support of Windows XP and the 32-bit software. In addition new options for virtualisation and emulation that are likely to appear in the mean time will also be monitored.

5.1. The proposal

As a result of these investigations, the project team proposed adding virtualisation to the current preservation workflow for software-based artworks at Tate. The current procedures will be maintained, with the steps related to virtualisation being added to it in the following way:

1. Description of the artwork.
2. Documentation of the hardware and software environments.
3. Identification of the artwork's significant properties and associated risks for long-term preservation, in discussion with the other stakeholders.
4. If software is bespoke, analysis of the function of the source code supplied, by someone other than the original programmer
5. Creation of exhibition copy. If any extra software must be added to the physical computer (e.g. libraries or drivers for

particular printers) then this should also be made a documented component of the artwork.

6. Creation of virtual machine using the physical to virtual process
7. Compare the significant properties in the physical and virtual machines. Specialist technical support is likely to be required to identify the less visible differences between the physical and virtual machines.
8. Add virtual machine to the virtualisation platform running the other Tate servers.
9. Create OVF file and test it on Virtual Box.
10. Add the OVF file to Tate's High Value Digital Asset (HVDA) storage system

All the new elements created must be tracked in the Tate's Collection Management System, "The Museum System" (TMS)[29], so a component with a unique identifier is created for the following elements:

- physical back-up machine,
- virtual machine on virtualisation platform
- OVF file on the HVDA system
- Individual software required, for example the operating system and particular libraries or drivers.

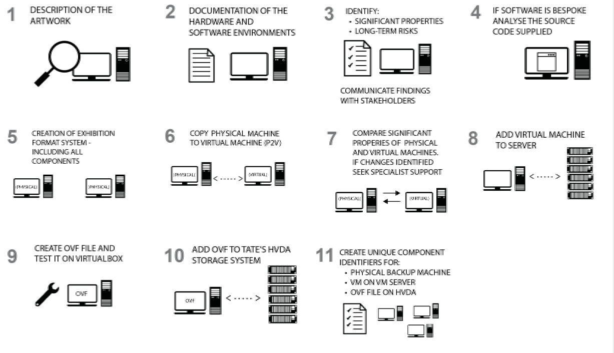

1 DESCRIPTION OF THE ARTWORK
2 DOCUMENTATION OF THE HARDWARE AND SOFTWARE ENVIRONMENTS
3 IDENTIFY:
• SIGNIFICANT PROPERTIES
• LONG-TERM RISKS
COMMUNICATE FINDINGS WITH STAKEHOLDERS
4 IF SOFTWARE IS BESPOKE ANALYSE THE SOURCE CODE SUPPLIED
5 CREATION OF EXHIBITION FORMAT SYSTEM - INCLUDING ALL COMPONENTS
PHYSICAL
PHYSICAL
6 COPY PHYSICAL MACHINE TO VIRTUAL MACHINE (P2V)
PHYSICAL
VIRTUAL
7 COMPARE SIGNIFICANT PROPERIES OF PHYSICAL AND VIRTUAL MACHINES. IF CHANGES IDENTIFIED SEEK SPECIALIST SUPPORT
PHYSICAL
VIRTUAL
8 ADD VIRTUAL MACHINE TO SERVER
9 CREATE OVF FILE AND TEST IT ON VIRTUALBOX
OVF
10 ADD OVF TO TATE'S HVDA STORAGE SYSTEM
OVF
11 CREATE UNIQUE COMPONENT IDENTIFIERS FOR:
• PHYSICAL BACKUP MACHINE
• VM ON VM SERVER
• OVF FILE ON HVDA

Figure 3-Workflow Diagram

5.2. Maintenance

General maintenance of the virtual machines would be carried out by the IS department, which maintains Tate's servers. This means the conservation department can rely on the pre-existing IS experience, infrastructure and maintenance protocols, avoiding the costs of creating a new infrastructure.

Conservation retains responsibility for testing the virtual machines at regular intervals and any major upgrades of the virtualisation platform. This testing involves comparing the significant properties identified at the beginning of the process and ensuring they remained the same, or within the agreed parameters.

6. CONCLUSION

The project identified virtualisation as a step towards a viable strategy for the preservation of our software-based artworks. Virtual machines will also in turn become obsolete. It may be that the virtual machines can be migrated, or that alternative strategies may be developed in the future to keep the software-based artwork operational. As with any digital object, preservation will need to involve the active monitoring and management of material to ensure that it remains accessible.

Virtualisation provides a complete environment within which the software runs, this enables comparison with our original systems; making it possible to check that they behave in the same way. An important aspect of this strategy is creating a virtualised version of the work, whilst the original can be confirmed as still running correctly.

Each software-based artwork is different, and it is an important aspect of the challenge of conservation to identify the significant properties of a particular artwork. Finding the best way to compare physical and virtual machines will also have to be decided on a case by case basis.

VMware would be our virtualisation platform of choice because it is already part of the infrastructure at Tate, and so procedures for maintenance have already been developed and put in place. Consequently conservation could utilise the expertise and resource already available in house, and did not need to create a parallel system. It is also a more mature tool and has been developed over a longer period.

VirtualBox was not completely discarded, as it is useful for creating exhibition copies, by running the original software in a new individual computer, not a server as required by the VMware tool used in Tate's IS department. It is also a good tool to test the compliance of the VMs created with VMware.

Given the advantage of emulation in running software independently of the underlying system architecture, and also the quick evolution in the tools available it is relevant to research the use of the tools available, namely QEMU.

The preferred method for virtualisation is to create a physical to virtual transfer, as this method captures all the contents of the computer, providing more information about the systems and processes used. Additional testing is required to establish the best method to carry out this transfer, as the tool used during the current tests installed an additional piece of software in the original machine. This is not considered best practice by the digital forensics community, as it introduces a change in the original system. We are therefore considering creating an initial disk image and then using the virtualisation software on the disk image. This method would avoid the need to make any changes in the original computer.

One of the main limitations identified was the virtualization of Apple Macintosh systems. As of this moment Apple Macintosh limits the running of Mac operating systems to Mac hardware, and circumventing this is possible, but raises a host of legal and copyright issues beyond the scope of this project.

Finally, one major outcome was the development of a scenario for the preservation of the software-based artworks, where we defined the steps we think will need to be taken. This scenario was made available to the partners in the Pericles project, and we are collaborating with them to develop useful tools that can help us with this workflow. One example is defining parameters for the Pericles Extraction Tool, which will help us automatically extract environment information, not only the usual system information but also software dependencies.

7. ACKNOWLEDGMENTS

We would like to thank Pip Laurenson and John McNeill in Tate’s Collection Care Research for their support of our project, the Conservation Department for allowing the time for this research and the other members of the Time-based Media Conservation team for all their help and support. This work was partially supported by the European Commission Seventh Framework Programme under Grant Agreement Number FP7-601138 PERICLES

8. REFERENCES

[1] Promoting and Enhancing Reuse of Information throughout the Content Lifecycle taking account of Evolving Semantics (PERICLES). http://pericles-project.eu

[2] Falcao, P. 2010. Developing a Risk Assessment Tool for the conservation of software-based artworks. MA Thesis, BFH-Hochschule der Kuenste Bern (HKB) https://www.academia.edu/6660777/Developing_a_Risk_Assessment_Tool_for_the_conservation_of_software-_based_artworks_MA-Thesis

[3] VMware. http://www.vmware.com/uk/products/vsphere/.

[4] VirtualBox. https://www.virtualbox.org/.

[5] DIOSCURI. http://dioscuri.sourceforge.net/faq.html#6

[6] QEMU. http://www.claunia.com/qemu/

[7] Jones, M. 2002. The Cedars Project, http://eprints.rclis.org/6045/1/article84a.pdf

[8] Rinehart, R. 2002. *Preserving the Rhizome ArtBase*. http://archive.rhizome.org/artbase/preserving-the-rhizome-artbase-richard-rinehart/

[9] Rinehart, R. 2002. *The Straw that Broke the Museum's Back? Collecting and Preserving Digital Media Art Works for the Next Century*. SWITCH: Online Journal of New Media. http://switch.sjsu.edu/web/v6n1/articlea.htm

[10] Rothenberg, J. 2003. *Renewing the Erl King*. http://www.bampfa.berkeley.edu/about/ErlKingReport.pdf

[11] National Library of the Netherlands. http://www.kb.nl/en/expertise/e-depot-and-digital-preservation/emulation/project-emulation-dioscuri

[12] Preservation and Long-term Access through Networked Services (PLANETS). http://www.planets-project.eu/http://www.planets-project.eu/.

[13] Keeping Emulation Environments Portable (KEEP). http://www.keep-project.eu/ezpub2/index.php

[14] http://www.digital-preservation.com

[15] Baden-Wuerttemberg Functional Long-Term Archiving and Access (bwFLA). http://bw-fla.uni-freiburg.de/

[16] Oberman, A. 2012. *ZKM Case Studies for the digital art conservation project* http://www02.zkm.de/digitalartconservation/index.php/en/case-studies.html

[17] Lurk, T., Enge J. 2008. Virtualisation as preservation measure. A contribution to the handling of born digital media art, in: *Archiving 2008. Final Program and Proceedings*, Ed.: Society for Imaging Science and Technology, Springfield, VA, Bern, 221-225

[18] PACKED. 2011. *Interview with Tabea Lurk and Juergen Enge*. http://www.scart.be/?q=en/content/interview-tabea-lurk-and-j%C3%BCrgen-enge

[19] Lorrain, E. 2013. PACKED Case Study report: Mondophrenetic (2000, Herman Asselberghs, Els Opsomer, Rony Vissers). http://www.scart.be/?q=en/content/case-study-report-mondophrenetic%E2%84%A2-2000-herman-asselberghs-els-opsomer-rony-vissers-0

[20] Matthews, B. mcIlwrath, B., Giaretta, D., Conway, E., 2008, *The Significant Properties of Software: A Study* . JISC http://www.jisc.ac.uk/media/documents/programmes/preservation/spsoftware_report_redacted.pdf

[21] Dipple, K. 2010. Final Report for the New Media Art Network on Authenticity and Performativity. Tate: London

[22] Paul, C. 2003. Digital Art. Thames & Hudson, London

[23] Transmediale, Berlin. http://www.transmediale.de

[24] Ars Electronica. http://www.aec.at

[25] Center for Art and Media Karlsruhe. http://on1.zkm.de/zkm/e/.

[26] Laurenson, P. "Old Media, New Media? Significant Difference and the Conservation of Software Based Art." In New Collecting: Exhibiting and Audiences after New Media Art, edited by Beryl Graham: Ashgate, 2014.

[27] http://www.tate.org.uk/art/artworks/martinat-mendoza-brutalism-stereo-reality-environment-3-t13251

[28] http://www.dmtf.org/standards/ovf

[29] The developers of TMS are Gallery Systems http://www.gallerysystems.com/tms

Risk Driven Selection of Preservation Activities for Increasing Sustainability of Open Source Systems and Workflows

Tomasz Miksa, Rudolf Mayer
Stephan Strodl, Andreas Rauber
SBA Research, Vienna, Austria

Ricardo Vieira, Goncalo Antunes
INESC-ID Information Systems Group
Lisbon, Portugal

ABSTRACT

The increasing demands faced by repository systems and the growing popularity of workflow systems introduces new risks, creating new challenges to the digital preservation community. The application of risk management practices to digital preservation is a way of managing the risks associated with the use of such systems and to optimize the application of digital preservation treatments to such risks. In this paper, we present results of a case study conducted on two use cases: a repository system and an automated workflow representing a typical digital preservation quality assessment process. We used a risk assessment approach to identify risks related not only to the technical but also organizational and legal aspects. We assigned controls which decrease their impact and explained how the digital preservation related controls can also improve the current functioning of the repository system and increase reproducibility of the workflows.

1. INTRODUCTION

In recent years, the digital preservation community has been investigating different ways of dealing with the preservation of static contents like scans of books or music recordings. Several solutions were proposed and successfully implemented. They range from frameworks and metadata vocabularies to distributed repository systems. Keeping up with the paradigm shift in science and the deluge of data [17], the digital preservation solutions are being enhanced to address the requirements of preserving complex objects like scientific data, workflows and processes. Nowadays, the digital preservation actions aim not only to safeguard the heritage to future generations but also to enhance the reproducibility of data-driven research.

There are several studies on repository systems which compare their functions and evaluate whether they address the needs of institutions in need of such [12]. Most of the systems are open source, which can be related to the fact that it is often assumed that open source solutions are considered to have higher preservability than their commercial counterparts. Such an assumption is also taken for the reproducibility of modern research, i.e. the scientific experiment is supposed to be reproducible when it is published under an open source license.

Yet, this assumption can be challenged. Being open source is not, by itself, a guarantee of the higher longevity of repository systems. At some point their preservation will be required, which can be a challenge due to the fact that these systems are more than databases which collect metadata and store the preserved objects. These systems are quite often distributed, benefiting from the integration of several external services which are provided by external entities. There is a potential risk that the functionality of the system can be severely affected when one of these services becomes unavailable or changes are made to its functionality.

In the worst case this may hinder the possibility of retrieving and presenting the preserved objects to the user. Similar threats are also affecting scientific workflows, which very often contain references to external services. Furthermore, workflows often require access to external libraries and software applications which need to be present during execution but are not explicitly defined in the workflow specification.

For this reason, we investigated potential threats to open source repository systems and workflows. A case study was conducted on two use cases: a repository system based on Fedora Commons, and a typical Taverna workflow used during preservation quality assessment. Both of them are available under the open source licenses. We used a risk assessment method based on ISO 31000:2009 and aligned with the TIMBUS preservation framework. The case study identified a wide spectrum of risks related not only to the technical but also organizational and legal aspects. We assigned controls which help to decrease their impact and detail the solutions delivered by the TIMBUS project that help to control the risks which are related, not only to digital preservation, but also to the current functioning of the repository system and reproducibility of modern research.

The paper is structured as follows. Section 2 presents the state of the art in risk management and explains the process preservation framework which guided our assessment. Moreover, changes in the web services that may affect both the repository system and the scientific workflows are discussed in this section. Section 3 describes two use cases on which the case study was performed. In Section 4, the approach used for risk assessment of both cases is presented and the results are discussed. Section 5 describes selected controls applied to both use cases. Finally, conclusions are presented in Section 6.

2. STATE OF THE ART

This section explains the risk management approach applied on the case study and provides an overview of available standards. It also places the risk management process within the process preservation framework. Finally, possible kinds of changes in web services are discussed.

2.1 Risk Management

Risk Management (RM) concerns the assessment and control of risks, with risk being defined as the combination of the likelihood of an event and its consequences [9]. Its ultimate goal is to manage the uncertainty associated with risks, either by mitigating risks with negative consequence on objectives or by taking advantage of risks with positive consequence on objectives [9].

Although different standards, methods and tools exist for targeting specific domains, ISO 31000:2009 [11] describes a generic and domain-independent framework for risk management, providing the underlying concepts and principles, along with a process. The risk management process defined by the standard is depicted in Figure 1.

The process starts by defining the internal and external context of the project. The external context might consist of a description of the regulatory environment of the project or any other element that might affect data management. The internal context includes defining all the elements of the project, i.e. its objectives, resources, data, processes, systems, among others that may be relevant to consider.

After establishing the context, the assessment of the risks based on the collected information is performed. It is composed of three different steps: (1) risk identification, where all relevant assets, vulnerabilities, events and risks are identified; (2) risk analysis, where the value of the assets, the exposure to vulnerabilities, the likelihood of events, the risk consequence, and ultimately the risk severity are estimated; and (3) risk evaluation, where the information produced in the two previous steps to check against risk criteria is evaluated, culminating in a decision on whether a specific risk is acceptable or tolerable. Depending on the context of the risk assessment, different risk assessment techniques can be applied to the process. The standard describes several of those techniques and their suitability for the different steps of the process.

These risk assessment steps result in the prioritization for risk treatment, with the identification of controls. If the controls are sufficient to lower the overall risk level into acceptable values, then a risk report is defined. All the steps of the process should be communicated to the interested parties for consultation and validation. Additionally, the process should be run continuously, with constant monitoring and review of the different steps, if necessary, so that the risk management is effective.

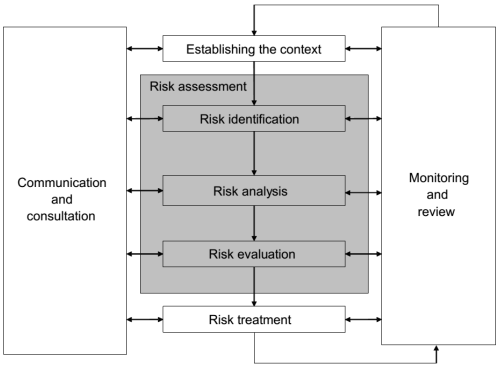


Figure 1: Risk management process according to ISO 31000:2009 [11]

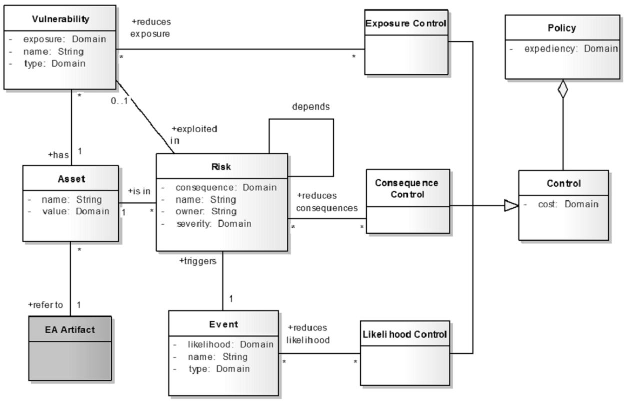


Figure 2: Risk concepts [2]

Digital preservation is one of the domains where risk management has been applied, as it is about recognizing that during its lifecycle, data is subject to risks that can affect their proper use and interpretation. Different works concerning risk management applied in this domain have been published, including the ISO 16363:2012 [10], that provides a risk management process for assessing the trustworthiness of digital repositories, and the Digital Repository Audit Method Based on Risk Assessment (DRAMBORA) [13], which also describes a process for assessing digital preservation repositories. Additionally, in [3], the authors identify a set of typical threats and vulnerabilities that can be mitigated using Digital Preservation techniques.

TIMBUS proposes a risk management-based approach to the preservation of business processes [15]. In that sense, digital preservation is seen as a risk treatment, with the clear interfaces existing between the TIMBUS processes and the risk management process. The risk management process adopted by TIMBUS follows the ISO 31000:2009 standard. A conceptual model is used along with the process, defining a set of risk management concepts, based on the work described in [2]. The model, which can be seen in Figure 2, is based on the ISO 31000:2009 family of standards, and was created to support sharing, reuse and processing of risk concepts. The model defines risk as "an effect of uncertainty

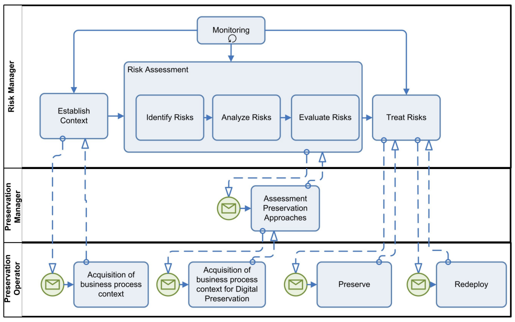


Figure 3: TIMBUS framework for process preservation, BPMN model

and is expressed by the combination of the likelihood of an event and its consequences when exploiting a vulnerability of an asset" [2]. Asset is defined as something (e.g. process, data, hardware, software, people) that has value to the project. A risk is expressed by a risk severity (or risk level) that is a combination of its consequence with the likelihood of the event triggering the risk. Finally controls are defined as actions that can be taken to mitigate risks. Controls can reduce the exposure of a vulnerability, reduce the likelihood of an event, reduce the risk consequence, transfer the risk and accept the risk. A risk policy represents a set of controls that were applied to mitigate the risks in a specific context.

2.2 Framework for Process Preservation

A process model for digitally preserving a process is described in detail in [16], and depicted in Figure 3. It is centred around the risk management approach detailed above, and can be divided into three phases: *plan*, *preserve* and *redeploy*.

The plan phase concerns the capture of the business process context. To this end, a context meta-model is used to systematically capture these aspects that are essential for its preservation and verification upon later re-execution [1]. This model is implemented in the form of an OWL ontology, which provides both generic core concepts and domain-specific concepts that are used for capturing information in specific domains. The generic concepts are based on ArchiMate [7], which provides a template to describe a business by around 30 different concepts on the business, application and technology layer. A number of domain-specific concepts dealing with *Software licenses* and *Patents*, *Software application dependencies*, and *Digital preservation meta-data*, among others, are provided.

Assessment of Preservation Approaches is responsible for the identification and evaluation of different preservation approaches (controls) for the process. Some specific controls will be discussed together with the use case description in Sections 4 and 5. Each approach is specified in a *Process Preservation Plan*, which also defines procedures for capturing the process data and later redeploying and verifying the process.

During the preservation phase, these controls are applied, and validation and verification data are captured from the source system for redeployment. The redeployment phase specifies the re-initiating of the preserved process in a new environment at some point in the future. Necessary adjustments to the target environment are performed, and finally, the process can be re-executed and the taken measurements can be validated.

2.3 Changes in services

According to the classification presented in [14], there are four ways in which a web service may change. In this section, those changes are discussed, as they may apply to these use cases and, additionally, because the general classification may apply to any kind of service (not only web services).

A web service can become unavailable. This will likely stop execution of the process, unless alternative paths and exception handling has been implemented for such a case. Reasons for unavailability can range from temporary technical problems, to bankruptcy of the service provider. Such situations are straight forward to detect, for instance, by using time-outs which would alert to the unavailability of the web service.

A web service can change its communication interface, not always jeopardising the full execution of the process. Such a situation may also be easily detected. It may require short pauses in the process execution until the changes are adopted into the process. Of course, in case of significant changes in the communication interface (e.g., switch from REST to WSDL), time needed for reconnecting the web service into the process may require more effort.

The functionality of a web service may change, which denotes that the outputs of the web service change, while the interface stays the same. Unlike the first two threats, this threat is hard to detect, as the process may not break, but instead will be delivering outputs which are not correct or are different from expected. Such a situation might be detected only much later and on a different level, e.g., when some general statistics regarding process performance are changed. Such a situation may occur for several reasons. One of the reasons may be the changes at the semantic level, e.g., switching the unit of measurement from inches to centimetre due to a server configuration change. Other possibilities are bug fixes in the underlying algorithm (which may introduce other bugs as well), or intentional changes in the functionality, e.g., faster but less accurate computational algorithms.

A web service may change its non-functional behaviour, which may not always stop the process from correct execution, but can occur temporarily and therefore be hard to notice. The examples of such cases could be different timing characteristics or delays, effects of buffering, etc. They also need to be detected, because there may be a threshold from which the web service cannot deliver its functionality properly, and therefore stop or alter execution of a dependent system.

3. USE CASES

In this section, an overview of two open source use cases is presented. The technical aspects that are relevant for the risk assessment presented later in this paper are explained. The first use case deals with a repository system, while the second use case is a typical digital preservation workflow.

Different as they may seem, the later analysis shows that they have much in common. We used these two cases in order to demonstrate the broad applicability of the TIMBUS preservation framework and the risk driven approach.

3.1 Fedora Commons Based Repository

This use case concerns a real installation of a repository system at a university. Due to the fact that the analysis described in this paper may reveal sensitive data, we cannot disclose any information which would allow its identification. Therefore in the remainder of the paper we will refer to the repository system used in the use case as the "repository system".

The repository system described here is a university-wide digital asset management system with long-term archiving functions. It offers the possibility of archiving valuable assets, together with normalised metadata and content formats, offering multilingual access. Students, researchers and co-operators with the proper authorisation can upload and link the objects which, among others, can be text, image, and audio files in multiple formats. Searching and browsing of the contents is possible without logging in.

The repository system is used in many ways. It holds scans of precious books and incunabula, which can be accessed through a book viewer module. Projects run in the different institutes and faculties of the university archive their collections of audio recordings or historical documents in the repository system. The importer module allows creation of virtual collections of different content types that can be grouped, archived and published together.

Implementation The system consists of two main components, the backend and the frontend. The backend is realized with the use of Fedora Commons[1], which is an open source system that allows for storing, managing, and accessing digital objects. It also provides modules for searching (GSearch: Fedora Generic Search Service) and interfaces for the exchange of metadata (OAI-PMH - Open Access Initiative Protocol for Metadata Harvesting). The web frontend is responsible for the presentation of contents, or editing of metadata. The frontend was developed at the university. The communication between these two components is realised through the use of XML interfaces (REST-Calls).

Content Transformation The Fedora repository holds local content in the form of digital objects. The frontend interacts with the Fedora repository through the Fedora API, as seen in the Figure 4. The backend may also interact with other systems to obtain the content stored on different servers (distributed content) or may use web services to get additional information about the contents or to perform data transformation (e.g., format conversion, video streaming). These services are of particular interest, because they may change in different ways and therefore alter the information and the content delivered to the end users of the repository system.

In order to understand why services are so crucial for digital objects in a Fedora based repository, it is important to understand the structure of a digital object. The digital object consists of four main parts: (1) a digital object identifier, (2) a descriptive part, including key metadata necessary to manage and discover the object and its relationships to other objects, (3) an item perspective which is the set of content or metadata items, and (4) a service perspective which provides methods for disseminating content.

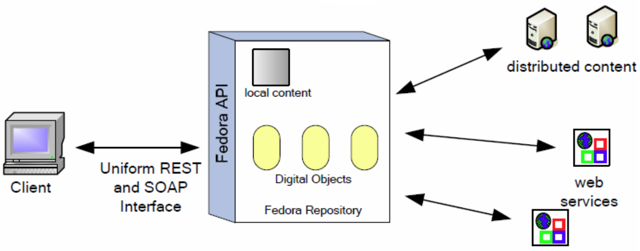


Figure 4: Backend (Fedora) as a mediator for services and content [6]

Such a structure allows creating an object which has all data provided as static or dynamic data. For example, in the case of static data, a scientific paper can be stored in the Fedora repository in three formats: HTML, PDF, TEX and all of them will be grouped under one digital object. No disseminators (services performing operations on content) will be used to produce the content, because the content will be provided at the moment of creation of the digital object. However, the same final (visible to the end user) result could be obtained with the use of disseminators. It would be possible to store, for example, only the TEX file and use web services to generate on demand a PDF or HTML version of the document. This second solution allows saving storage space in the repository, but introduces dependency on the services (disseminators). Such dependencies are unavoidable in case of interactive contents like interactive art, games, computer programs, which need a special environment to render these artefacts.

Key functionalities We will discuss in the following section the key functionalities provided by the repository system. They have high impact on the content presented to the user and introduce other dependencies which may need to be considered during the risk analysis.

The Image Converter is used every time users access a web page with a summary about a digital document. For example, if the user browses through a collection of PDFs and opens one of them then they are presented with a preview of the first page of the paper which is a PNG file generated from the first page of the original file. This is achieved with the use of ImageMagick[2] and the corresponding Perl module which needs to be installed in the operating system underlying the repository system. If a different version of ImageMagick is used, it may happen that the conversion may result in a different output. Therefore, all of the dependencies of the repository system need to be documented carefully in order to be able to reproduce the same rendering.

The repository system also uses a streaming server which is

[1] http://www.fedora-commons.org

[2] http://www.imagemagick.org

run by the IT department of the university and is not a part of the repository system itself. When a video is accessed through the repository system there is a check if the video is already available at the streaming server. If it is, then the video is played to the user; otherwise the video has to be decoded by the server and then presented to the user. The video is displayed in a web browser window. In both cases, the streaming server has to be available. If it is not, or if it has changed (e.g. different codec library installed), the user may be presented with different rendering or, in the worst case, with no rendering at all.

3.2 Quality Assessment Workflow

Workflows have become popular as a means for specifying and automating computational experiments [5]. They serve a dual function: first, as detailed documentation of the executed process (i. e. the input sources and processing steps taken for the computation of a certain data item), and second, as re-usable, executable artefacts for data-intensive analysis. Using a workflow, a process is defined as a series of analysis steps which specify the flow of data between them.

We investigated a number of workflows published by third parties, many of them in the domain of digital preservation, such as workflows for file characterisation or format migration. The workflow that we will utilise as case study in this paper deals with duplicate detection in the book digitisation domain [8]. When scanning books, a software searches for errors in the scanned images. Due to the time lag of the error detection, errors are usually detected only when the scanning process has already scanned several more pages since the event occurred.

In such a case, the scanning process goes back to the erroneous page and restarts from there, thus re-scanning the erroneous and subsequently scanned pages. As the initial set of scans is not deleted, this leads to a set of duplicate scanned pages. The purpose of the duplicate detection is to perform a quality assurance of the document collection before the ingest into a repository system. The workflow specifically runs a duplication detection, and evaluates the performance of that duplication detection, with the help of a manually created ground truth that correctly identifies duplicate pages. Knowing the performance of the duplication detector is important to evaluate whether relying on the software solution only is sufficient, or a manual quality control step is needed in addition.

The workflow is authored in the Taverna workflow engine, and depicted in Figure 5. The actual duplicate detection is done via the *matchbox* application, developed as part of the SCAPE project. Matchbox (visible as the step "matchbox" in Figure 5) is implemented in Python, and called from the workflow via an external tool invocator, i.e. a system call. More specifically, the matchbox application is not available locally on the machine that executes the Taverna workflow, but is accessed as a remote service, via an SSH (secure shell) connection, on a different server. The output of the Matchbox algorithm is parsed in the "parse_matchbox_stdout" step, and the resulting matches as well as the log output of the application, are available as process outputs.

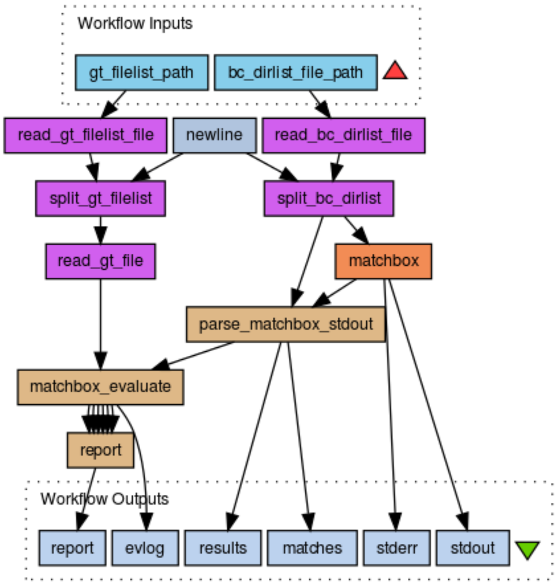


Figure 5: Duplicate Detection workflow, authored in the Taverna Workflow engine

In order to evaluate the performance of the duplicate detection, reported matches are compared against a previously provided ground truth (passed via the process input "gt_filelist_path"), which contains the above mentioned true information on which pages are actual duplicates. This evaluation is implemented as a Java Beanshell script "matchbox_evaluate", which calls functions from a Java library provided in a JAR file to this processor step, and provides the workflow output "report", which combines correctly/incorrectly identified duplicates, missed duplicates, and measurements, such as precision, recall, and F-measure.

While workflows are a step towards longevity and preservation of process executions, they themselves are not sufficient, as the execution environment, and external dependencies, are not properly addressed. In the use case example, an interesting aspect undermining that fact is the call to the Matchbox application, which is not performed via a local system call, but via an SSH connection on a remote server. Not only does that introduce a dependency on an external system, this call is also protected by the standard secure shell authentication mechanism, requiring a user name and password. To run the workflow, one thus additionally needs these credentials. Furthermore, the dependency on this external system means that the functioning of the workflow depends entirely on the functioning of that service.

Another interesting aspect is the step matchbox_evaluate, which, as mentioned above, is a Beanshell script that extensively uses an externally provided Java library for most of its functionality. This dependency to the Java library is declared in the workflow, but the actual library is not part of the workflow definition file, and thus has to be preserved separately. Another difficulty is the complex structure of the output of matchbox. All output information is returned in one text file, with a custom-defined format structuring the information on which pictures are identified as duplicates. Firstly, there is no documentation on the exact output format available, and in addition, the format was slightly changing throughout the different versions of development of "matchbox".

The matchbox_evaluate component, which processes the output to do the evaluation, was developed by a different organisation than the duplicate detection itself, and the development of these two components was not always synchronised. Thus care has to be taken that the versions used of these two components are compatible to each other. Configuration management can help with this, but as there is not much information available on when the remote service would change its interface or structure and format of the returned result (such as at least notifications of a change), this issue is not easily resolved.

4. RISK ASSESSMENT

In this section, we explain how we followed the TIMBUS preservation framework described in Section 2.2 and depicted in Figure 3. We also present the results for both of the use cases.

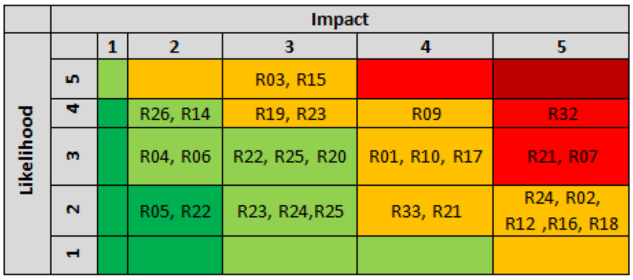
4.1 Performing the assessment

Following the risk assesment process depicted in Figure 1 and described in Section 2.1, our first step was the identification of assets, events, risks and potential consequences. We utilized a combination of following techniques to collect this information:

- Checklist - list of risks previously defined, resulting from previous assessments with similar objectives. In this case, we used domain-specific lists, namely DRAMBORA and TRAC (Trustworthy Repositories Audit & Certification: Criteria and Checklist).
- System analysis - system/workflow documentation including its model and direct investigation of the running system/workflow. This technique involves analysing several processes performed by the system/workflow from different perspective (business view, infrastructure view)
- Brainstorming - it had the objective of identifying risks that were not detected from the checklists and system analysis.
- Semi-structured interviews - the risk assessment team met with the system operators to conduct individual interviews. They were asked a set of questions, encouraging them to look at a situation from a different perspective and thus identify new risks.
- Legal risks were analysed in accordance to the national and international legal documents by legal experts.

In the following steps, we analysed the risks and events. For each of the events we assigned the likelihood using a range of 5 values: very low, low, medium, high, very high. We used the same scale for assessing the consequences of the risks. Having done this, we created a risk matrix (see Figure 6). The dimensions of the matrix are likelihood and impact. By putting each risk into the cell that corresponds to its likelihood and impact, an overview of the severity of the risks was obtained.

The colours of the cells represent the risk level classes according to the established risk criteria, which represent the associated range of severity. This helped us to understand which risks need special attention when designing controls in the next step. The controls were designed for each of the risks identified and were applied to treat the risk, thus decreasing its severity. Naturally, the process of risk assessment needs to be periodically repeated in order to confirm that there are no new risks and if the controls are mitigating the risks efficiently.



		Impact				
Likelihood		1	2	3	4	5
	5			R03, R15		
	4		R26, R14	R19, R23	R09	R32
	3		R04, R06	R22, R25, R20	R01, R10, R17	R21, R07
	2		R05, R22	R23, R24,R25	R33, R21	R24, R02, R12 ,R16, R18
	1					

Figure 6: Risk Matrix for the repository case

4.2 Results

We applied the risk assessment process described in the above section to both of the use cases. In this section we present an overview of the results and explain the selection of controls.

Fedora Commons Based Repository We identified 29 events and 19 risks which constituted in total 46 pairs of events triggering the risks. We associated them with 5 different asset types, namely: organization, repository functionality, data stored in repository, repository system software. Such a wide variety of assets shows that the risk assessment focused not only on the risks related to digital preservation and software availability, but also on the risks related to the organizational context and legal issues. Table 1 shows some examples of pairs of events and risks identified, together with the assets they concern.

According to the risk assessment process we formulated controls for each of the risks and events. The results of this analysis were presented to the management of the repository and are going to be a base for the discussion on improvements to the repository system and its broad context. In the remainder of this section we would like to focus on a subset of controls which can be introduced using concepts and tools delivered by the TIMBUS project. Thus we demonstrate how the solutions from the digital preservation domain can mitigate a wide range of risks and not only those which are directly related to digital preservation.

Table 4.2 presents a list of TIMBUS related controls which can be used to minimize the likelihood of events or consequence of risks. It also depicts how their values change after control application. One can notice that some of the controls appear more than once in the controls column, e.g. *Context Model Instantiation*. Thus by applying one control we benefit from mitigating multiples risks. Furthermore, in the case of *Context Model Instantiation* we are controlling at one time events related to the typical digital preserva-

Table 1: Subset of assets, events and risks for the repository case

Asset	Event	Risk
Organization	Change of business model	Financial loss due to change of business model
Repository Functionality	Internal or external attacks	Functionality fault due to internal or external attack
Repository Functionality	Loss of expert knowledge	Functionality faults due to loss of expert knowledge
Organization	User's illicit activities	Reputation loss due to ilicit use of repository from user
Repository Software	Software faults	Software unavailability due to software faults
Data	Changes to content model	Loss of data integrity due to changes in data model

tion problems like *Environment changes*, but also business related risks like *Loss of expert knowledge* or *Changes in organizational structure*. In Section 5 we show how two of these controls were implemented, i.e. *Context Model Instantiation* and *External dependencies monitoring.*

Quality Assurance Workflow The subset of the risks identified for this use case is given in Table 4.2. It has to be noted that this use case has much fewer social and other contextual aspects to consider compared to the repository use case. This is due to the workflow itself being a mostly technical artefact, and the original environment where the workflow was executed being unknown and thus not a part of the use case. Thus, most risks concern technical aspects.

One important group of risks concerns the externally provided services, i.e. the duplicate detection algorithm. This service may be hosted outside of the organisation running the workflow, and outside of their control. Furthermore, the service is not following any protocol such as a WSDL web service, the communication, expected parameters and expected return values are not explicit in the workflow definition. Finally, the service requires an authentication, for which the user name and password are not kept along with the workflow definition.

Another group of risks is concerned with the library used in the workflow to process the results from the duplicate detection. This library can become unavailable, as it is not packed together with the workflow definition, or can have a fault, or can be incompatible with the version of the external service. The workflow engine itself is also a risk, as it may become unavailable, or incompatible with the operating system the workflow is deployed on. Finally, knowledge about the workflow setup, execution and interpretation is very often implicit, tacit knowledge of the owner of the workflow. If that person is not available anymore (e.g. due to changing jobs), it might not be possible to run the workflow anymore.

A subset of the identified controls for the workflow use case is given in Table 4.2. Some controls are the same or similar to the ones identified for the repository case; in general, similar observations as in the repository use case hold true: some of the controls appear more than once, thus mitigating more than one risk at the same time.

5. CONTROLS

In this section, we describe the selected controls and their application to the use cases in detail. The description explains how the control works, and in what way the risks or events are controlled. We selected the controls which address the entities on a high level of likelihood/consequence.

5.1 External dependencies monitoring

External dependencies monitoring is aimed at identifying the types of changes described in Section 2.3. Detecting this type of changes will not have an impact on the likelihood of the events happening, but it can help to reduce the consequence. This is on the one hand due to being able to detect a change earlier than when detecting it by a process triggered by the system itself. Potentially, we are able to detect and issue a fix to the issue before the problem surfaces in any process execution. Further, having monitored the service, we might have data available that allows us to more quickly identify the specific problem with the service, thus being able to find a solution for it quicker. External dependencies monitoring is applied in both use cases, as they both rely heavily on them.

Regarding the repository system, all of the services used are currently hosted within the infrastructure of the university. However, not all of them are under direct control of the repository system support team. Furthermore, due to the constant development of the repository system and provision of new services, it is likely that some of the services may be provided by external partners. The repository system can use any kind of web service regardless of its location. Such flexibility may cause potential threats. For example, when a service is down, many functions of the repository system depending on it will become unavailable or at least have their functionality limited. Furthermore, changes in the implementation of the external service may be unnoticed, but may impact the system. For this reason, we decided to monitor external services for their availability and changes and have response scenarios prepared in advance to mitigate the consequences of the service change.

We implemented the control using the Web Service Monitoring Framework (WSMF)[14]. The WSMF allows intercepting traffic communication between the system and the analysed web service. During standard operation of the service the data intercepted is stored as ground truth data. It is later used for validation of the service. We periodically sent the requests gathered in the ground truth data to the monitored web service and compared the responses with the ground truth data responses. On this basis we can detect whether the behaviour of the web services is changed. Figure 7 presents the application that allows performing these actions. For now we are able to detect changes in the Image Converter module of the repository system. The WSMF can also be applied to monitor web services responsible for conversion of content model by disseminators.

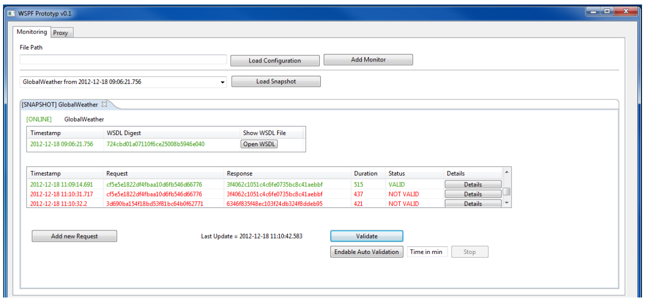
As shown in Table 4.2, the control *External dependencies monitoring* decreases the consequence of *Functionality faults*

Table 2: Subset of controls for the Repository use case

Control name	Control type	New Value	Old Value	Controlled Entity
List of users with administration rights	Likelihood	medium	high	Modification using administration rights
Context Model Instantiation	Consequence	medium	high	Changes in organizational structure
Context Model Instantiation	Consequence	medium	high	Loss of Expert Knowledge
External dependencies monitoring	Consequence	medium	high	Functionality faults
Group policies	Likelihood	medium	high	Modification using administration rights
Context Model (Infrastructure View)	Likelihood	low	medium	Environment changes
Mock-ups of services	Consequence	medium	high	Functionality fault
Preservation of system and data	Consequence	low	medium	Shortcomings in semantic understandability
Software escrow	Consequence	medium	high	Functionality fault
Substiution of missing components	Consequence	low	high	Functionality fault

Table 3: Risks identified for the Workflow use case

Event	Risk
Authentication failure	External service unavailability due to authentication failure
Correct Library version not found	Workflow execution failure due to library dependency unavailability
Data files not available	Workflow execution failure due to unavailability of data dependencies
Library faults	Workflow execution failure due to library dependency faults
Library unavailability	Workflow execution failure due to library dependency unavailability
Loss or lack of documentation	Shortcomings in semantic understandability due to loss or lack of documentation
External Service faults	Workflow execution failure due to external service dependency fault
External Service unavailability	Workflow execution failure due to external service dependency unavailability
Workflow engine faults	Workflow execution failure due to workflow engine fault
Workflow engine unavailable	Workflow execution failure due to workflow engine unavailability
Workflow executed on unsupported OS	Workflow execution failure due to unsupported operating system



GlobalWeather from 2012-12-18 09:06:21.756

[SNAPSHOT] GlobalWeather

[ONLINE] GlobalWeather

Timestamp	WSDL Digest	Show WSDL File
2012-12-18 09:06:21.756	724cbd01a07110f6ce25008b5946e040	Open WSDL

Timestamp	Request	Response	Duration	Status	Details
2012-12-18 11:09:14.691	cf5e5e1822df4fbaa10d6fb546d66776	3f4062c1051c4c6fe0735bc8c41aebbf	515	VALID	Details
2012-12-18 11:10:31.717	cf5e5e1822df4fbaa10d6fb546d66776	3f4062c1051c4c6fe0735bc8c41aebbf	437	NOT VALID	Details
2012-12-18 11:10:32.2	3d690ba154f18bd53f81bc64b0f62771	6346f835f48ec103f24db324f8ddeb95	421	NOT VALID	Details

Last Update = 2012-12-18 11:10:42.583

Figure 7: Web Service Monitoring Framework control panel [14]

from the high to the medium likelihood. This is because any potential changes influencing the functionality of the repository system are quickly noticed and instant preventive actions can be taken.

For the Workflow use case, the application is very similar - we can also apply the WSMF to the external service used, and are thus able to detect any changes at an earlier stage. Thus, we can reduce the consequences of external services becoming unavailable. According to Table 4.2 the consequence of *External service unavailability* is reduced from high to low.

5.2 Context Model Instantiation

The context model, introduced in [1] and described in Section 2.2 gives a comprehensive picture of the environment the process is embedded in. This allows for a documentation of the process steps, the actors, and their connection and dependency towards the technological infrastructure that provides the execution platform in a comprehensive and formal manner. Its formal representation enables reasoning, and checking for compliance.

The context model is an important control addressing a number of risks identified in our use cases. In some of these risks, having a context model that covers only the infrastructure aspects of our systems is enough, as these risks primarily deal with dependencies and conflicts between software applications. For some other risks, especially those regarding the knowledge on how the process is executed, a full-scale model that also covers application and business aspects is required.

Regarding the repository use case, even if the services are available locally the content presented to the user still may be altered in comparison to what was projected previously. Such a situation may occur when a user accesses some content (e.g., a video) which is rendered with different algorithms (e.g. different video codecs) and therefore may have a different look and feel of the digital object. In most cases, this is not a big issue for daily use, but in terms of digital preservation and documenting the significant properties of the digital object correctly for preservation purposes it is of great significance. Therefore when preserving a repository system, the knowledge about all of the elements impacting the final representation of the object have to be documented.

Also the dependencies of the repository system need careful documentation, because they also may affect the final result presented to the user. The repository system depends on many Perl modules and new implementations of modules may introduce changes in the behaviour of the system. Hence, it is crucial to maintain information about the software dependencies of the system in order to be able to recreate the same look and feel, as well as behaviour at any time in the future.

On-going development of the system, such as changes and enhancements of metadata schemas in order to enable Repository system to archive contents from various scientific disci-

Table 4: Controls for the Workflow use case

Control name	Control type	New Value	Old Value	Controlled Entity
Substitution of missing components	Consequence	low	high	Application dependency fault
External dependencies monitoring	Consequence	medium	high	Application license expired
Substitution of expired components	Consequence	low	high	Application license expired
Context Model (Infrastructure View)	Likelihood	low	medium	Application or Library incompatibility
Context Model (Infrastructure View)	Likelihood	low	medium	Application unavailability
Storing credentials for external services	Likelihood	medium	high	Authentication failure
Context Model (Infrastructure View)	Likelihood	low	medium	Correct Library version not found
External dependencies monitoring	Consequence	medium	high	Data files not available
Archiving and Preservation of data	Consequence	medium	high	Data files not available
Substitution of faulty components	Consequence	low	high	Library faults
Context Model (Infrastructure View)	Likelihood	low	medium	Library unavailability
Context Model Instantiation	Consequence	medium	high	Loss or lack of documentation
External dependencies monitoring	Consequence	medium	high	External Service faults
Mock-ups of services	Consequence	medium	high	External Service unavailability
Software escrow	Consequence	medium	high	External Service unavailability
Context Model (Infrastructure View)	Likelihood	low	medium	Workflow executed on unsupported OS

plines, creates another preservation requirement. For example, some digital objects may have been described through use of a metadata schema, which was later modified by adding new classifications and voluntary fields. However, it may happen that this new information cannot be added to the existing elements. These elements may then appear to a future user as corrupted, because the user may think that some of the metadata is missing despite the fact that the schema (the newer one) enforces its existence. The problem becomes even more complex when the concepts used in different versions of the schema are redefined and change their meanings. In order to prevent incorrect reasoning and wrong conclusions about the objects, it is essential to preserve the original versions of the metadata schemas and couple them with objects using them. All of these can be described in the Context Model. Due to the non disclosure agreements we are not allowed to present the example of the Context Model for the repository case.

Concerning the open source workflow use case, the technical part of the context model is an effective control regarding dependency and incompatibility risks. With the concepts provided by the meta-model, we can formally capture the dependencies between the application and library components used in the system. This helps when identifying issues that could be caused by changing versions of certain parts of the system setup. Furthermore, by having the full instantiation of the context model, it becomes clear what sequence of steps is needed to be carried out in the process, and how each step is supported by certain parts of the infrastructure. Also, existence of external services become clear, and their impact to certain parts of the process is explicit. The data flow between the steps is formally defined, which helps in understanding how the data is processed.

A simplified version of a corresponding instance of the TIMBUS Context Model is depicted in Figure 8. The model depicts the external system that is called via SSH. It also shows the third-party library that is needed for the matchbox algorithm evaluation.

5.3 Application Substitution

An application is usually utilised to manipulate or render a digital object. By replacing (substituting) the original application interpreting the digital object the functionality of this application is emulated. This is an effective control to mitigate risks that can stem from faulty or incompatible software applications, libraries and components utilised in the system. By replacing them with another component that provides equivalent behaviour, but does not exhibit the risks, we successfully mitigated that risk by application of emulation.

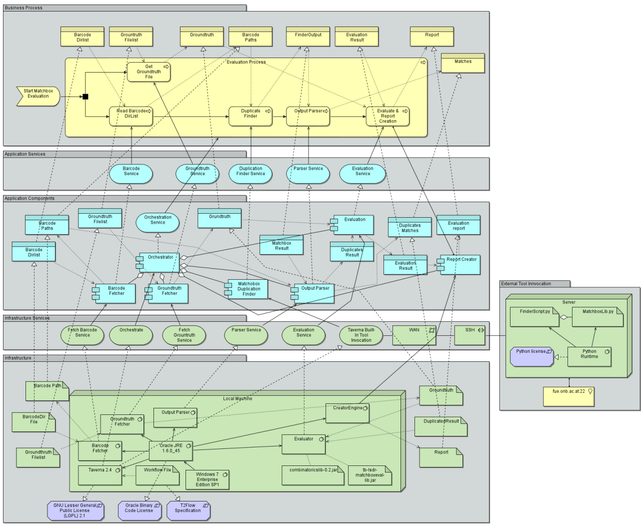


Figure 8: Context Model of the Duplicate Detection workflow

As part of the TIMBUS project, we developed a service that allows for automatic identification of potential alternative software implementations, and thus application emulation. The service is built around knowledge bases obtained from linked data sources such as Freebase[3] [4], as well as software packages as they are present often in Linux operating systems, where *virtual packages* provide a categorisation of packages that provide the same functionality. The service then operates on a representation of the system, authored by using the context model, and proposes potential replacements, that in turn should be analysed by a digital preservation expert for their usefulness and feasibility.

[3] http://www.freebase.com/

This approach can be used in the case of the repository system to decrease the consequence of *Functionality fault* risk (see Table 4.2). For example if the Tomcat application server that is a container for most of the repository backend is obsolete and loses the community support, it can be replaced with a compatible one, like Jetty, which may not have this problem. Moreover, multiple Java and Perl libraries may also need to be substituted. One of the potential reasons could be low security of the component, then such a vulnerable library may be replaced with a recommended alternative. This shows again that the digital preservation tools can also ease day to day maintenance of the system.

6. CONCLUSIONS

This paper describes the results of a case study conducted on two use cases: a repository system and an automated workflow representing typical digital preservation quality assessment processes. We followed the TIMBUS preservation framework and risk assessment process defined by ISO 31000:2009 to identify potential risks and their impact on the sustainability of systems and workflows.

The case study revealed a wide range of risks affecting not only the technical aspects of the cases but also organizational aspects. First and foremost, it confirmed the concerns that the repository systems may need to undergo several digital preservation actions. Hence, there should be more attention to this problem within the digital preservation community and the contents of the repositories are not the only thing we need to worry about. Furthermore, the preservability of both systems and workflows is endangered due to a high dependence on external services and insufficient documentation of their dependencies.

Using tools developed within the TIMBUS project we demonstrated how these risks can be substantially mitigated. We used the external dependencies monitoring, context model instantiation and application emulation as controls to achieve this aim.

ACKNOWLEDGMENTS

This research was co-funded by COMET K1, FFG - Austrian Research Promotion Agency, by the FCT - Fundação para a Ciência e a Tecnologia, under project PEstOE/EEI/LA0021/2013, and by the European Commission under the IST Programme of the 7th FP for RTD - Project ICT 269940/TIMBUS.

7. REFERENCES

[1] G. Antunes, M. Bakhshandeh, R. Mayer, J. Borbinha, and A. Caetano. Using ontologies for enterprise architecture analysis. In *Proceedings of the 8th Trends in Enterprise Architecture Research Workshop (TEAR), in conjunction with the 17th IEEE International EDOC Conference*, Vancouver, Canada, September 9-13 2013.

[2] J. Barateiro. *A Risk Management Framework Applied to Digital Preservation*. PhD thesis, Universidade Técnica de Lisboa, Instituto Superior Técnico, 2012.

[3] J. Barateiro, G. Antunes, F. Freitas, and J. Borbinha. Designing digital preservation solutions: a risk management based approach. *The International Journal of Digital Curation*, 5:4–17, 2010.

[4] K. Bollacker, C. Evans, P. Paritosh, T. Sturge, and J. Taylor. Freebase: A collaboratively created graph database for structuring human knowledge. In *Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data*, SIGMOD '08, pages 1247–1250, New York, NY, USA, 2008. ACM.

[5] E. Deelman, D. Gannon, M. Shields, and I. Taylor. Workflows and e-science: An overview of workflow system features and capabilities. *Future Generation Computer Systems*, 25(5):528–540, May 2009.

[6] Fedora Commons Community. Fedora commons tutorial 2: Getting started: Creating fedora objects using the content model architecture. Technical report, 2007.

[7] T. O. Group. *ArchiMate 2.0 Specification*. Van Haren Publishing, 2012.

[8] R. Huber-Moerk, A. Schindler, and S. Schlarb. Duplicate detection for quality assurance of document image collections. In *Proceedings of the 9th International Conference on Digital Preservation (IPres2012)*, Toronto, Canada, October 1-5 2012.

[9] ISO. *ISO Guide 73:2009 – Risk management – Vocabulary*. International Organization for Standardization, 2009.

[10] ISO. *ISO 16363:2012 – Space data and information transfer systems – Audit and certification of trustworthy digital repositories*. International Organization for Standardization, 2012.

[11] ISO/FDIS. *ISO/FDIS 31000:2009 – Risk management – Principles and guidelines*. International Organization for Standardization, 2009.

[12] Y. Li and M. Banach. Institutional repositories and digital preservation: Assessing current practices at research libraries. *D-Lib Magazine*, 17(5/6), 2011.

[13] A. McHugh, R. Ruusalepp, S. Ross, and H. Hofman. The digital repository audit method based on risk assessment (DRAMBORA). In Digital Curation Center and Digital Presevation Europe, 2007.

[14] T. Miksa, R. Mayer, and A. Rauber. Ensuring sustainability of web services dependent processes. *International Journal of Computational Science and Engineering (IJCSE)*, 2014. Accepted for publication.

[15] S. Strodl, D. Draws, G. Antunes, and A. Rauber. Business process preservation: How to capture, document & evaluate? In *Proceedings of the 9th International Conference on Preservation of Digital Objects*, Toronto, Canada, 1–5 October 2012.

[16] S. Strodl, R. Mayer, G. Antunes, D. Draws, and A. Rauber. Digital preservation of a process and its application to e-science experiments. In *Proceedings of the 10th International Conference on Preservation of Digital Objects*, Lisbon, Portugal, September 2013.

[17] M. Van der Graaf and L. Waaijers. *A Surfboard for Riding the Wave. Towards a four country action programme on research data. A Knowledge Exchange Report.* 2011.

Management and Orchestration of Distributed Data Sources to Simplify Access to Emulation-as-a-Service

Thomas Liebetraut and Klaus Rechert
Albert-Ludwigs University Freiburg
Hermann-Herder-Str. 10
79104 Freiburg i. B., Germany
{firstname.lastname}@rz.uni-freiburg.de

ABSTRACT

Emulation-as-a-Service makes emulation widely available for non-experts and thus, emulation could prove valuable as a tool in digital preservation workflows. Providing these emulation services to access preserved and archived digital objects poses further challenges to data management. Digital artifacts are usually stored and maintained in dedicated repositories and object owners want to or are required to stay in control over their intellectual property.

In this paper we propose a distributed storage and data access model that ensures that the user stays in control over his digital objects by simultaneously providing efficient data transport and support for (space) efficient management of user modifications. Finally, a mechanism for orchestration of both storage and emulation services to re-enact a single pre-defined setup is presented.

General Terms

Infrastructure

Keywords

Emulation as a Service, Distributed Data, Framework, Cloud Computing

1. INTRODUCTION

Emulation of legacy computer systems is technically challenging and requires computing power as well as specialized knowledge about computing technology. These challenges pose a hurdle to non-technical users of emulation services that want to preserve and access digital objects like interactive art or legacy software. The goal of the Emulation-as-a-Service (EaaS) [10] framework is providing emulation services to these non-technical users like memory institutions or owners of digital object collections.

To implement the EaaS service model and make it usable for preservation purposes, a certain modularization and division of duties is required. Therefore, the EaaS framework is divided into the actual emulation service provided by the service provider, archives storing and maintaining digital objects provided by their respective owners, and modular workflows to access and interact with the digital object. While providing and maintaining emulation components requires highly specialized knowledge and will probably always be done by specialized service providers, the archive component is designed to be provided by different institutions.

Libraries and owners of collections of digital objects, may want to or are even required to stay in control over their intellectual creations, making it necessary to keep these digital objects in a separate archive. Consequently, there are potentially many decentrally organized archives that are operated independently from each other. They all may have different requirements on how to maintain and create the archived data and there may be little or no coordination between different archive providers. In some cases, users may choose not to use a public archive or storage service and create their own micro-archives that suit their specific needs. Some may only exist over the course of a single session. In such a decentralized structure, archives and emulation services may appear and disappear as well as digital objects may be relocated to other archives. But also object owners or users may decide to switch to a different EaaS provider. For this, we propose a comprehensive set of interfaces and metadata to orchestrate an EaaS service and coordinate access to multiple heterogeneous archives in a unified way.

An EaaS service provider may opt to provide various ready-made emulated computer environments, so-called base images with operating systems (OS) and drivers already installed and configured, sharing the costs of maintenance and technical expertise to create these environments. Instances of emulated environments, i.e. an installed and configured OS plus software stack on a virtual disk image, may reach up to hundreds of GB in file size. Even with currently available network bandwidth, copying a full environment to an EaaS Cloud service is inefficient and impairs the user experience. In addition, users may need to change, customize or personalize environments. Hence, user modifications need to be tracked and stored for subsequent usage. Therefore, we propose a distributed storage and data access interface that (1) ensures that the user stays in control over his digital objects, (2) provides efficient data transport even with limited bandwidth and (3) supports efficient management of user modifications.

2. RELATED WORK

The concept of legacy platform emulation is closely tied to the development of computer systems and is well established as a tool to bridge a technological gap [7]. Recently, emulation has evolved as a tool for preservation of complex digital assets [12, 16, 17]. Furthermore, emulation setups have been formalized to assess authenticity and performance [4], and specific aspects of simulation of individual technical components such as CRT screen simulation have been addressed [3, 14].

While these works have greatly promoted emulation in a scientific context as well as the professional use of emulation in digital preservation, many of these aspects have to be orchestrated and implemented individually for each purpose. For instance, emulation has been used to provide access to a large collection of legacy CD-ROMs [2, 18]. Furthermore, requirements and workflows have been developed for preparing ready-made environments to render certain digital artifacts [11]. To enable several institutions to make use of and potentially contribute to the collection, the digital objects were made available through a distributed filesystem and required a specific emulator setup on the user's site.

The KEEP project [1] addressed this problem by networked provisioning of various complex emulator setups [5, 8]. While a networked approach reduces technical and organizational hurdles on the client's side significantly, it still requires technical expertise and manual tasks carried out by the user. Furthermore, data management, especially maintenance of specific environments has not been addressed yet.

A more community-centered approach is the Olive platform [2], which is specifically designed to allow collaboration of different curators on a Cloud-based library. Olive also uses local emulation using a thin client approach to run virtual machines, but it uses its own protocol to stream data necessary to execute the virtual machine over the network. Modifications to a virtual machine, for example, newly installed software, can be transferred back to the archive, making derivatives of digital objects possible [13]. With our proposed data management approach, we split generic computer and software environments, potentially ready-made for emulation purposes, from highly specific user adaptions and user data. This way, the object owner remains in complete control of both how and by whom the objects are accessed as well as how and by whom the objects are stored and maintained but still benefits from cost-effective shared maintenance of common components.

3. REQUIREMENTS & ARCHITECTURE

Emulation-as-a-Service is built as a distributed architecture that separates the different tasks required to re-enact a single digital object. This separation allows for every component to be maintained by respective specialists. Basically, the EaaS model is divided into the emulation service itself that handles the emulation task, and archives that provide digital objects. While common objects like operating systems and software can be shared in federated storage archives, e.g. to share maintenance costs, digital objects preserved at a memory institution remain in the full control of these institutions (cf. Fig. 1).

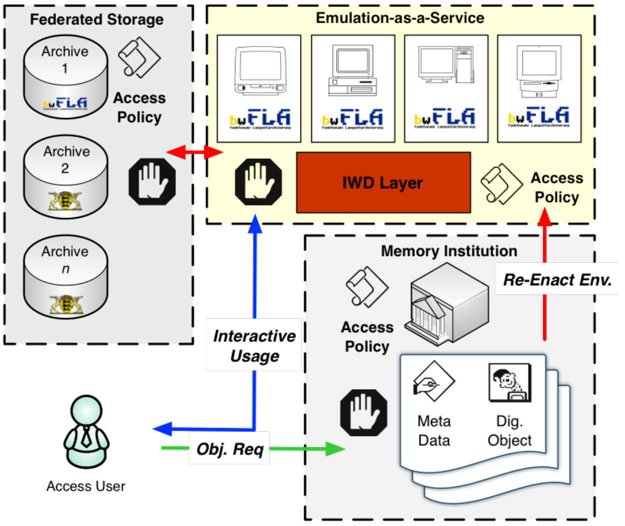


Figure 1: The EaaS distributed architecture with service provider and different archive providers.

In order to provide a public EaaS service model, an abstract description of how these different entities are to be coordinated is necessary. This description can then be used by an EaaS service's emulation components to bring together all the necessary bits and pieces to enable interactive user access to complex digital objects. Hence, the emulation component should not make any assumptions on the structure of the archive storing requested digital objects. Similarly, the archive or respectively its description must not assume a specific implementation of the emulation site. Finally, for emulation-based preservation of digital objects technical meta-data should be an abstract description of how to re-enact a specific computer environment that does not depend on a particular emulator software that will ultimately face the same digital obsolescence problem like all digital objects and technology.

3.1 Emulation Environment

To allow an individual computer environment to be replicable in the future, an abstract description of such a computer environment is required that is independent from emulator-specific configuration or its implementation details. Therefore, we introduce a comprehensive and abstract description of a computer system, the *emulation environment*. This technical metadata describes a computer environment to an extent that an emulation component can use it to reproduce the original environment. It includes the hardware architecture (platform) to be emulated as well as all devices that are optional to that platform (disk drives, sound cards, input devices, etc.).

These device descriptions might depend on external resources or assets like firmware ROM code or disk images that consist of binary data. For instance, an operating system, software and other digital objects are provided on emulated media

[1] Keeping Emulation Environments Portable, http://www.keep-project.eu/, last retrieved 2/1/2014
[2] https://olivearchive.org/

types such as virtual hard disks or CD-ROMs. After this data is created or retrieved from actual media or hardware and is preserved on a bit level, it has to be made available to the EaaS framework.

The data archives that provide preserved digital objects are not necessarily part of a specific EaaS service but can be provided by different data-centers or institutions. This means that all the digital objects required by an emulation environment may not be directly available for the EaaS service provider. Therefore, the data objects that are required by an emulation environment are referenced using *data bindings*.

These bindings reference the digital object using an URL that identifies the object's location or using a persistent identifier. Each binding is identified and accessed by the emulation component using an identifier unique within the emulation environment. To the emulation component these bindings are independent of the actual data location, access policy and other properties. This is achieved by the use of special data connectors that hide the complexity of actually accessing the digital object's URL and provide a simple, file-like access method to the data. Certain details of this access can still be specified by the emulation environment, though, for instance enforcing a specific transport protocol. While the data in an archive has to be read-only to guarantee long-term preservation constraints and to support efficient concurrent access, bindings always have to be writable from an emulator point of view. Technical restrictions in the emulated operating systems and saving customizations to the environment requires modifications to be tracked and stored for subsequent usage. To make the emulation environment metadata useful for archival and preservation purposes, it can also be extended with descriptive metadata like environment title, authoring information and creation dates. Similarly, the description may also contain information about what software is installed in the environment or which digital objects can be accessed.

The emulation environment is the basic building block to orchestrate the different components of the EaaS framework. It allows for separation of the emulation component and the archive and makes it possible to view emulation environments as a real document that can be referred to and be collaborated on. Changes made to an emulation environment can be ingested back into an archive which makes them again available as a new, derived environment.

3.2 Persistent Identification

While the emulation component heavily relies on the availability of data, the origin if this data does not matter. In the case of archives provided outside of the EaaS service provider, using static references to an archive to link the emulation environment with associated data is not feasible and would complicate migration to other EaaS or storage providers. Especially when implementing the archive component using dynamic Cloud storage solutions that can be allocated on-demand, referencing data by its network location (i.e. IP or host name) is not applicable as data can move to another host and may only be available for a limited time at the specified network location.

To solve this problem it makes more sense to ignore image locations altogether and refer to data using a unique and persistent identifier (PI) such as Uniform Resource Name (URN), Digital Object Identifier (DOI), or The Handle System (HDL) [1]. If the archive moves to another host or some digital objects move to another archive (or are distributed among many archives), the same PI can be used to resolve all available image locations, allowing load-balancing and dynamic allocation of resources in the cloud.

3.3 Persistent User Sessions

Once objects are stored in an archive and an appropriate environment has been created to access these objects, the environment should be immutable and cannot be modified except explicitly by an administrational interface. This guarantees that a memory institution's digital assets are unaltered by the EaaS service and remain available in the future. It also allows efficient concurrent access handling without the need to implement a complex and possibly expensive data and session management to avoid interfering with other users' sessions.

This immutability, however, is not easy to handle for most emulated environments. Just booting the operating system may change an environment in unpredictable ways. When the emulated software writes parts of this data and reads it again, however, it probably expects this data to represent its modifications. Also, users that want to interact with the environment must be able to change or customize it permanently. Therefore, data connectors have to provide write access for the emulation service while they cannot write the data back directly to the serving archive.

4. IMPLEMENTATION

The outlined requirements are used to orchestrate several components required to make digital objects in auxiliary archives accessible by an EaaS service instance. Individual data bindings that represent a single digital object are connected to by using *data connectors* on the EaaS site that are configured by the binding specification in the emulation environment. They can then be referenced by URLs of the form `binding://identifier`, e.g. to define a hard disk's data. Data connectors provide a generic interface between the archives and the actual emulation software to access heterogeneous data sources. They implement the network transport protocol, handle network connectivity and provide all the input and output operations that are common for a standard local file, like reading, writing and random access. Optionally, they also provide methods to authenticate the current user session to the archive if this is necessary to access protected digital objects. Different data connectors can be provided to support different network transport and authentication protocols in order to access different memory institutions' archives.

This concept requires some support from the archive to make archived objects accessible from the EaaS framework. Usually, digital objects from archives are not accessible directly as a single bit-copy of the original medium. Elaborate housekeeping information and further metadata is usually stored alongside the original object. To allow data connectors to access the individual digital object over the network, an archive server component has to be deployed at the memory institution's site that translates the internal data structures used to archive the digital object to a network protocol suitable for accessing these objects. This archive component hides the complexities of bookkeeping and accessing preserved objects while granting or restricting access to individual objects. Consequently, the archive component can be highly specific to the needs and structure of the archive that are usually determined by the archiving institution. At the same time, it enables the EaaS service to access the raw data of individual digital objects in a unified way.

The distributed nature of this approach requires an efficient network transport of data to allow for immediate data access and usability. However, digital objects stored in archives can be quite large in size. When representing a hard disk image, the installed operating system, together with installed software, can easily grow up to several GB in size. Even with today's network bandwidths, copying these digital objects in full to the EaaS service may take minutes and derogates the user experience. While the archived amount of data is usually large, the data that is actually accessed frequently can be very small. In a typical emulator scenario, read access to virtual hard disk images is block-aligned and only very few blocks are read by the emulated system [15]. Transferring only these blocks instead of the whole disk image file is typically more efficient, especially for larger files.

Therefore, the network transport protocol has to support random data access and sparse reads without the need for actually copying the whole data file. While direct file access provides these features if a digital object is locally available to the EaaS service, it is not applicable in the general case of separate emulation and archive servers. Special-purpose network file systems like NFS (Network File System) or SMB (Server Message Block) provide file-like access to remotely exported files over the network. They, however, require a complex setup in the host operating system of both, the emulation service itself and the archive servers at the memory institutions. Additionally, this setup has to be done for every archive server that has to be available to an individual emulation component.

In contrast, the Network Block Device (NBD) [6] protocol provides a simple client/server architecture that allows direct access to single digital objects as well as random access to the data stream within these objects. Furthermore, it can be completely implemented and run without administrational privileges on the host operating system and has a very simple software design that does not require a complex infrastructure on the archive servers.

4.1 Handle It!

In order to access digital objects, the emulation environment needs to reference these objects in the emulation environment. Individual objects are identified in the NBD server by using unique export names. Consequently, a URL schema of the form `nbd:<hostname>:<port>:exportname=<name>` can be used to declare the network location of an individual digital object.

While this NBD URL schema directly identifies the digital object and the archive where the digital object can be found, the data references are bound to the actual network location. In a long-term preservation scenario, where emulation environments, once curated, should last longer than a single computer system that acts as the NBD server, this approach has obvious drawbacks. Furthermore, the Cloud structure of EaaS allows for interchanging any component that participates in the preservation effort, thus allowing for load-balancing and fail-safety. This advantage of distributed systems is offset by static, hostname-bound references.

Therefore, the Handle System is used as persistent object identifier throughout our reference implementation to identify resources. The Handle System provides a complete technological framework to deal with these identifiers (or "Handles" (HDL) in the Handle System) and constitutes a federated infrastructure that allows the resolution of individual Handles using decentralized Handle Services. Each institution that wants to participate in the Handle System is assigned a prefix and can host a Handle Service. Handles are then resolved by a central resolver by forwarding requests to these services according to the Handle's prefix. As the Handle System, as a sole technological provider, does not pose any strict requirements to the data associated with Handles, this system is used as a PI technology.

Each Handle consists of a set of typed records that the Handle server has to return upon request. While there are some predefined record types like "URL" or "EMAIL", individual Handle Services are able and encouraged to define their own record types that fit their needs. As currently the only information required in bwFLA is the actual network location, the URL type is used to encode the actual NBD URL. Because there can be more than one record of the same type in a Handle, several of these URLs can point to different archive providers or provide different transport types. Handles are then referred to in the emulation environment using URLs of the form `hdl:11270/61fecaebea36...` where `11270` is the prefix registered to the bwFLA project and the following string an arbitrary identifier.

The Handle Service resolving Handles for the bwFLA prefix is installed locally on one of the network nodes that run the bwFLA software, but is available globally. While it makes sense for owners of digital objects to make use of similar Handle Services to consistently refer to their objects independently from the archives that host the object, it is not expected that this is an inherent part of the EaaS infrastructure. As Handles are used throughout the emulation environments to identify data, the Handle Service has to be independent of a specific EaaS provider in order to preserve these emulation environments and possibly migrate them to different EaaS providers. This can be achieved either by each owner of digital objects to register his own Handle prefix, or to provide a institutionalized service similar to the DOI foundation that is more suitable for the needs of data expected by the EaaS framework.

4.2 Persistent User Sessions

The concept of interacting with re-enacted environment is an important part of the EaaS framework. Both, base systems provided by the service provider that curators can build their own environment on and users that interact with the final environment require modifications to an existing environment. Only saving these modifications and making them accessible to others makes sharing of resources possible, reducing maintenance costs. At the same time it opens new possibilities for community-based curation efforts to allow contemporary witnesses to fine-tune and improve the user experience of digital objects like art or software [9].

A single EaaS instance not only consists of the digital objects themselves but also includes the emulation environment as orchestration and management metadata. Modifications to this metadata can easily be handled because the emulation environment can simply be copied due to its small file size. If the user attaches new drives or otherwise modifies the metadata, a new emulation environment can be created that includes the new hardware as well as the configuration of the base system. In most cases, however, the hardware environment does not change but the data on hard disks or other drives does. For example, installing software or configuring the software environment result in modifications to the underlying data. Also, just booting the operating system may change an environment in unpredictable ways and users that want to interact with the environment may change certain aspects of it. When the emulated software writes parts of this data and reads it again, it expects this data to represent its modifications.

As digital objects are not to be modified directly in the archive, a mechanism to store modifications locally at the EC while reading unchanged data from the archive has to be implemented. Such a transparent write mechanism can be achieved using a copy-on-write access strategy. While NBD allows for arbitrary parts of the data to be read upon request, not requiring any data to be provided locally, data that is written through the data connector is tracked and stored in a local data structure. If a read operation requests a part of data that is already in this data structure, the previously changed version of the data should be returned to the emulation component. Similarly, parts of data that are not in this data structure were never modified and must be read from the original archive server. Over time, a running user session has its own local version of the data, but only those parts of data that were written are actually copied.

We used the qcow2 container format[3], part of the QEMU project, to keep track of local changes to the digital object. Besides supporting copy-on-write, it features an open documentation as well as a widely used and tested reference implementation with a comprehensive API, the QEMU Block Driver. The qcow2 format allows to store all changed data blocks and the respective metadata for tracking these changes in a single file. To define where the original blocks (before copy-on-write) can be found, a *backing file* definition is used. QEMU's Block Driver API provides a continuous view on this qcow2 container, transparently choosing either the backing file or the copy-on-write data structures as source.

This mechanism allows modifications of data to be stored separately and independent from the original digital object during an EaaS user session, allowing to keep every digital object in its original state as it was preserved. Once the session has finished, these changes can be retrieved from the emulation component and used to create a new, derived

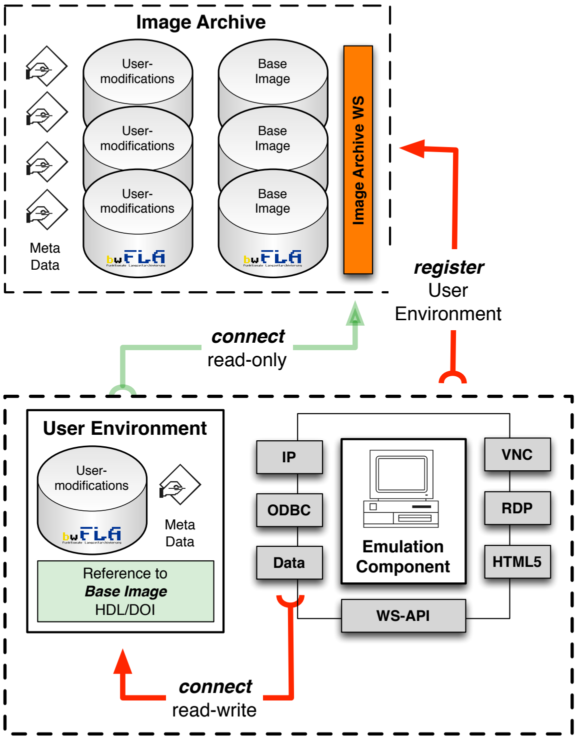


Figure 2: Data access workflow for derived environments. The eser environment exists only at the EaaS service provider until it is explicitly registered at the archive (if allowed).

data object (cf. Fig. 2). As any Block Driver format is allowed in the backing file of a qcow2 container, the backing file can also be a qcow2 container again. This allows "chaining" a series of modifications as copy-on-write files that only contain the actually modified data. This greatly facilitates efficient storage of derived environments as a single qcow2 container can directly be used in a binding without having to combine the original data and the modifications to a consolidated stream of data. However, this makes such bindings rely not only on the availability of the qcow2 container with the modifications, but also on the original data the qcow2 container refers to. Therefore, consolidation is still possible and directly supported by the tools that QEMU provides to handle qcow2 files.

Alternatively, a filesystem-based approach like UnionFS [4] could be used to track, store and maintain changes made to a system. These unification filesystems "stack" several modification layers on top of each other. While a filesystem-based approach offers convenient tools to track individual files, the metadata required to reconstruct these changes is implementation specific. Using a simple, block-oriented approach of maintaining a virtual disk's differential changes has some advantages in a digital preservation scenario, due to its simple meta-data structure. The result of changed blocks are a simple entries in a block mapping table (c.f. Listing 1 which defines which file the data should be read from. This simple

[3] The QCOW2 Image Format, `https://people.gnome.org/~markmc/qcow-image-format.html`, last access 8/15/14.

[4] A Stackable Unification File System, `http://unionfs.filesystems.org/`

representation allows a manual reconstruction, even if the original implementation is not available anymore.

Listing 1: An excerpt from the block mapping table used in qcow2.

```
Offset       Length    Mapped to  File
0            0x10000   0x270000   derived.qcow2
0x10000      0x10000   0x60000    base.qcow2
0x10000000   0x10000   0xab0000   base.qcow2
0x20000000   0x210000  0x50000    derived.qcow2
0x20210000   0x800000  0x2b0000   base.qcow2
0x30000000   0x10000   0xac0000   base.qcow2
0x3ffe0000   0x20000   0x80000    base.qcow2
```

Once the data modifications and the changed emulation environment are retrieved after a session, both can be stored again in an archive to make this derived environment available. If there is no efficient transparent write support and a full copy is used instead, the changed copy can be used directly. In case of a copy-on-write approach, only those chunks of data that actually were changed by the user have to be retrieved. These, however, reference and remain dependent on the original, unmodified digital object. It can then be accessed like any other archived environment.

4.3 Collection containers

Sometimes it is useful to archive several individual data objects combined in a single container. For example, when a software is distributed on more than one installation medium, all the images belong to the same software with each single one of them useless without all the other. To make this collection one single digital object, they can all be tied together into a container format, e.g. a UDF image or a tar archive. To refer to this new digital object and access individual images from it, the data connectors in our reference implementation support a mechanism to access the contents of containers.

To determine whether a digital object is a container, the data references can be used. If only the `binding://name` form is used, the digital object is accessed directly. As soon as a reference of the form `binding://name/subobject` is used to make use of a sub-object, the binding `name` is used as a container, requiring the use of the "collection connector" to access the data. To avoid implementing the NBD access protocol twice, this collection connector can be used on top of the NBD connector.

4.4 Example

Listing 2 shows an example emulation environment from our reference implementation describing an IBM OS/2 system. Apart from some management information like the title or an ID, it identifies the system architecture (line 4) and includes a drive specification (lines 9–17). The drive specification tells the EC about the virtual disk interface to use (line 11) and all necessary bus information. To refer to the data contained in the virtual hard disk, a special URI scheme referring to a binding is used instead of the actual location of the virtual hard disk image (line 10). This binding (lines 29–33) is then defined in terms of an HDL reference with automatic transport protocol negotiation in case the HDL resolves to more than one transport method (line 31). Finally, the binding also selects the copy-on-write access method (line 32) instead of a full copy, essentially enforcing a failure if none of the archives support random-seek read access. A second drive (lines 19–27) together with another binding (lines 35-38) demonstrates how the binding mechanism can be used for larger collections of floppy images for which it makes sense to archive them in one single container (e.g. as a tar archive or a UDF image). Subcomponents of this container can be accessed directly in the emulation environment with the EC providing an appropriate data connector to unpack this container.

Using this information, the EaaS framework can determine an EC suitable for emulating the requested system architecture (x86 PC). The EC then instantiates a suitable emulator configuration and connects to all defined bindings by the mechanisms described above. Additional environment configuration like an attached CD-ROM containing some digital artifact could be added by simply adding another `<drive>` element and choosing the correct PI for the CD-ROM. Likewise, the binding-mechanism also makes it possible to declare ROM-images or similar data.

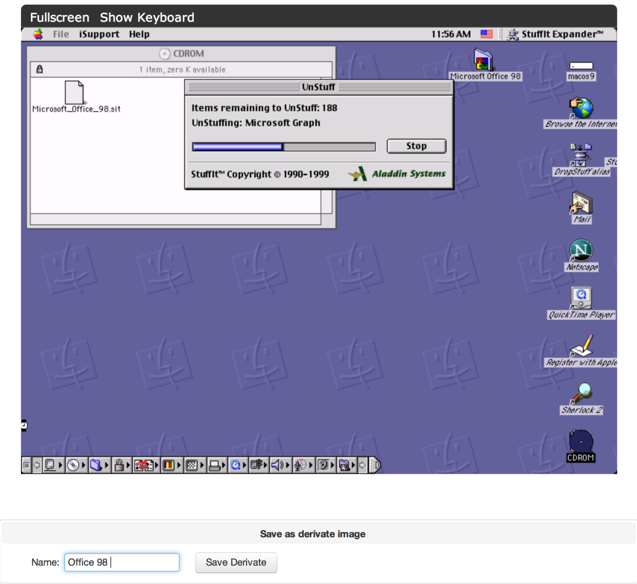
Listing 2: An example emulation environment configuration.

```
 1 <emuEnvironment xmlns="EmuEnvironment">
 2   <uuid>2016</uuid>
 3   <title>IBM OS/2 2.11</title>
 4   <arch>i386</arch>
 5   <description>
 6     ...
 7   </description>
 8
 9   <drive>
10     <url>binding://system_hdd</url>
11     <iface>ide</iface>
12     <bus>0</bus>
13     <unit>0</unit>
14     <type>disk</type>
15     <boot>true</boot>
16     <plugged>true</plugged>
17   </drive>
18
19   <drive>
20     <url>binding://floppys/disk1.img</url>
21     <iface>fdc</iface>
22     <bus>0</bus>
23     <unit>0</unit>
24     <type>floppy</type>
25     <boot>false</boot>
26     <plugged>true</plugged>
27   </drive>
28
29   <binding id="system_hdd">
30     <url>hdl:11270/0ecd47a3...</url>
31     <transport>auto</transport>
32     <access>cow</access>
33   </binding>
34
35   <binding id="floppys">
36     <url>hdl:11270/c41d0444...</url>
37     <transport>auto</transport>
38     <access>cow</access>
39   </binding>
40 </emuEnvironment>
```

For digital preservation purposes, it is often not sufficient to have this functional description of an environment. If any component of the emulation environment (especially the data referred to by bindings) is lost, the original purpose of the environment can no longer be determined. Therefore, the `<description>` element in the emulation environment contains a behavioral description of the emulated computer system like operating system, installed software, special configuration and customization this software underwent and other curation information. Using this archival information, a curator could, if he had access to all single original software components, re-create the complete environment.

5. USE-CASES AND EXAMPLES

To provide a better understanding of the EaaS image-archive interfaces and prototypical implementation, the following three use-cases demonstrate how the current implementation can be used in practical scenarios. An obvious scenario is the creation of so called derivatives of emulated computer systems, i.e. specifically adapted system environments suitable to render a specific object or to be used in a specific context. In a similar scenario a data object is injected into the environment which is then modified for later access, i.e. installation of a viewer application and adding the object to the autostart folder. Finally, an existing hard disk image (e.g. an image of a real machine's hard disk) is ingested into the system. This scenario requires, besides the technical adaption of the hardware environment suitable to be run in an emulator, private files are to be removed before public access.



Items remaining to UnStuff: 188
UnStuffing: Microsoft Graph

Save as derivate image

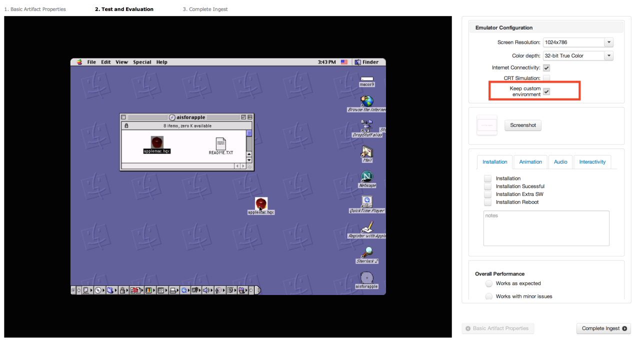
Name: Office 98

Save Derivate

Figure 3: Installing uploaded software package and creating a derivative environment.

5.1 Derivatives – Tailored Runtime Environments

Typically, an EaaS provider provides a set of ready-made environments, so-called base images. These images contain a basic OS installation which has been configured to be run on a certain emulated platform. Depending on the user's requirements, additional software and/or configuration may be required, e.g. the installation of certain software frameworks, text processing or image manipulation software. To do so, the user is able to upload a software installation package, which is then injected into the emulated environment, e.g. as CD-ROM or DVD medium. Once the software is installed, the modified environment can be saved and made accessible for object rendering or similar purposes (cf. Fig. 3).



Emulator Configuration

Screen Resolution: 1024x786

Color depth: 32-bit True Color

Internet Connectivity: ☑

CRT Simulation: □

Keep custom environment ☑

□ Installation
□ Installation Sucessful
□ Installation Extra SW
□ Installation Reboot

notes

Overall Performance

○ Works as expected

○ Works with minor issues

Figure 4: Ingest of CD-ROM art. Object is copied to the compouter's desktop and added as "autostart" object.

5.2 Object-specific Customization

In case of complex CD-ROM objects with rich multimedia content from the 90s and early 2000s such as encyclopedias and teaching software, typically a custom viewer application has to be installed to be able to render its content. For these objects, an already prepared environment (installed software, autostart of the application (cf. Fig. ??)) would be useful and would surely improve the user experience during access as "implicit" knowledge on using an outdated environment is not required anymore to make use of the object. Since the number of archived media is large, duplicating for instance a Microsoft Windows environment for every one of them would add a few GB of data to each object. Usually, neither the object's information content nor the current or expected user demand justify these extra costs. Using derivatives of base images, however, only a few MB are required for each customized environment since only changed parts of the virtual image are to be stored for each object. In the case of the aforementioned collection of multimedia CD-ROMs, the derivate size varies between 348kB and 54MB.

5.3 Authenticity vs. Redaction

Another scenario of increasing importance is the preservation complete user system like the personal computer of Villem Flusser in the Villem Flusser Archive [5]. Such complete system environments usually can be achieved by creating a hard disk image of the existing computer and use this image as the virtual hard disk for EaaS. Such hard disk images can, however, contain personal data of the computer's owner. While EaaS aims at providing interactive access to complete software environments, it is impossible to restrict this "interactiveness", e.g. to forbid access to a certain directory directly from the user interface. Instead, our approach to this problem is to create a derivative work with all the personal data being stripped from the system. This allows users with sufficient access permissions (e.g. family

[5] Villem Flusser Archive, `http://www.flusser-archive.org/`

or close friends) to access the original system including personal data, while the general public only sees a computer with all the personal data removed. The redacted version of the disk image is inextricably linked to the original image, such that any action of the redaction process can be audited.

6. CONCLUSION & OUTLOOK

The presented architecture and implementation provides means to connect an external archive to an EaaS infrastructure and to curate its objects using emulation-based preservation workflows. It provides a functional view on both, data and the hardware configuration of a computer system instead of specifying a direct network location or hardware model, both of which may be meaningless in the far future.

At the same time, the EaaS service allows to make preserved environments accessible to a broad audience and provides a community-centered curation approach in which changes made by individual users to improve the authenticity of an environment can easily be made available to the rest of the community without losing the original version of the environment. This also makes it possible to track improvements and understand how computer systems and software works, allowing for a better restoration process in the future.

The interfaces and architecture presented in this paper also provide several features to overcome common problems in a distributed network. First, large digital objects can be accessed efficiently over the network. First, digital objects can now be efficiently accessed over the network. Together with a location-independent PI to reference data, this allows for a complete separation of the archive and the emulation services, also on an organizational level. New digital objects do not need to be registered at the EaaS service provider and the emulation service does not require direct access to the archive's storage backend in order to re-enact a single object's behavior and utility. Digital objects can rather be used directly after making them available using either their NBD network location directly, or, preferably, after they have been registered at some PI service. As this service is usually not dependent on the implementation of a specific EaaS framework, this is a much more versatile approach. This also leads to the possibility of quickly adding new archives to the system without having to coordinate with the EaaS service provider. The pure archive component can easily be implemented on any platform and does not rely on specific features to be available. The reference implementation should be able to run on any POSIX compatible system with network access without any modifications. Therefore, using EaaS and the proposed data management concept, object owners are able to present their objects (interactively) without actually releasing the environment and, more importantly, the intellectual property to the user. This is a required feature for digital art and similar digital assets: to provide access to an almost unlimited amount of users in order to unfold its potential impact on today's society, e.g. to use and interact with a piece of digital art, without anyone being able to copy it. The owner remains in control of the object and is able to restrict access any time simply by restricting access to their archive.

Second, the use of a copy-on-write mechanism together with a transport protocol that allows fragmented access improves the user experience. Instead of having to wait for a full copy of the digital objects to be made, only minimal amounts of data have to be transferred in order to make the environment usable immediately after the initialization. Furthermore, it allows a community-centered curation approach in which changes made by individual users to improve the authenticity of an environment can easily be made available to the rest of the community without losing the original version of the environment. This also makes it possible to track improvements and understand how computer systems and software works, allowing for a better restoration process in the future.

Finally, a more structured emulation environment allows for a more future-proof approach to emulation-based preservation. The emulation environment separates the functional description of a hardware system and the archival metadata required to understand the system. Each of them can be exchanged independently from each other, either using a different approach to describe the hardware in a possible future EaaS solution, or using a different preservation metadata that describes how the environment was built and preserved and how it can be used.

7. REFERENCES

[1] Arms, W. Y., May 2001. Uniform resource names: Handles, purls, and digital object identifiers. *Commun. ACM 44*, 5 (May 2001), 68–.

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

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

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

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

[6] Machek, P., 1997. Network block device. http://atrey.karlin.mff.cuni.cz/~Davel/nbd/nbd.htm.

[7] Marsland, T., and Demco, G., 1978. A case study of computer emulation. *Canadian Journal of Operational Research and Information Processing 2* (1978), 16.

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

[9] Rechert, K., Espenschied, D., Valizada, I., Liebetraut,

T., Russler, N., and von Suchodoletz, D., 2013. An architecture for community-based curation and presentation of complex digital objects. In *Digital Libraries: Social Media and Community Networks, ICADL 2013* (2013), Springer, pp. 103–112.

[10] Rechert, K., Valizada, I., von Suchodoletz, D., and Latocha, J., 2012. bwFLA – A Functional Approach to Digital Preservation. *PIK – Praxis der Informationsverarbeitung und Kommunikation 35*, 4 (2012), 259–267.

[11] Reichherzer, T., and Brown, G., june 2006. Quantifying software requirements for supporting archived office documents using emulation. In *Digital Libraries, 2006. JCDL '06. Proceedings of the 6th ACM/IEEE-CS Joint Conference on* (june 2006), pp. 86–94.

[12] Rothenberg, J., 1995. Ensuring the longevity of digital information. *Scientific American 272*, 1 (1995), 42–47.

[13] Satyanarayanan, M., Bala, V., St. Clair, G., and Linke, E., 2011. Collaborating with executable content across space and time. *7th International Conference on Collaborative Computing: Networking, Applications and Worksharing (CollaborateCom)*, October (2011), 528–537.

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

[15] Tang, C., 2011. Fvd: a high-performance virtual machine image format for cloud. In *Proceedings of the 2011 USENIX conference on USENIX annual technical conference* (Berkeley, CA, USA, 2011), USENIXATC'11, USENIX Association, pp. 18–24.

[16] van der Hoeven, J., van Diessen, R., and van der Meer, K., 2005. Development of a universal virtual computer (uvc) for long-term preservation of digital objects. *Journal of Information Science 31*, 3 (2005), 196–208.

[17] van der Hoeven, J., and van Wijngaarden, H., 2005. Modular emulation as a viable preservation strategy. In *Proceedings of the 9th European Conference on Research and Advanced Technology for Digital Libraries* (Berlin, Heidelberg, 2005), ECDL'05, Springer-Verlag, pp. 485–486.

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

Epimenides: Interoperability Reasoning for Digital Preservation

Yannis Kargakis
Institute of Computer Science, FORTH-ICS
Greece
kargakis@ics.forth.gr

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

René van Horik
Data Archiving and Networked Services, DANS
Netherlands
rene.van.horik@dans.knaw.nl

ABSTRACT

This paper presents `Epimenides`, a system that implements a novel interoperability dependency reasoning approach for assisting digital preservation activities. A distinctive feature is that it can model also *converters* and *emulators*, and the adopted modelling approach enables the *automatic reasoning* needed for reducing the human effort required for checking (and monitoring) whether a task on a digital object (digital collection in general) is performable. Finally, the paper describes (in the form of scenarios) concrete preservation activities of a research data archive (DANS) and elaborates on how `Epimenides` could be used and the benefits that would bring.

Keywords

Conversion/Emulation, Dependency Management, Automated Reasoning, Case Study

1. INTRODUCTION

Can we achieve interoperability without necessarily having to rely on standards, but by combining existing software? This question is complex and difficult to answer, therefore the adoption of (or at least assistance from) an automated reasoning approach is beneficial. This is the objective of the migration and emulation-aware dependency reasoning that was presented in [12] (more in [6]). This paper describes the system `Epimenides`, the first system that implements this automated reasoning approach for digital preservation. The paper also elaborates on how it can be used in practice by a research data archive such as DANS (Data Archiving and Networked Services, NL).

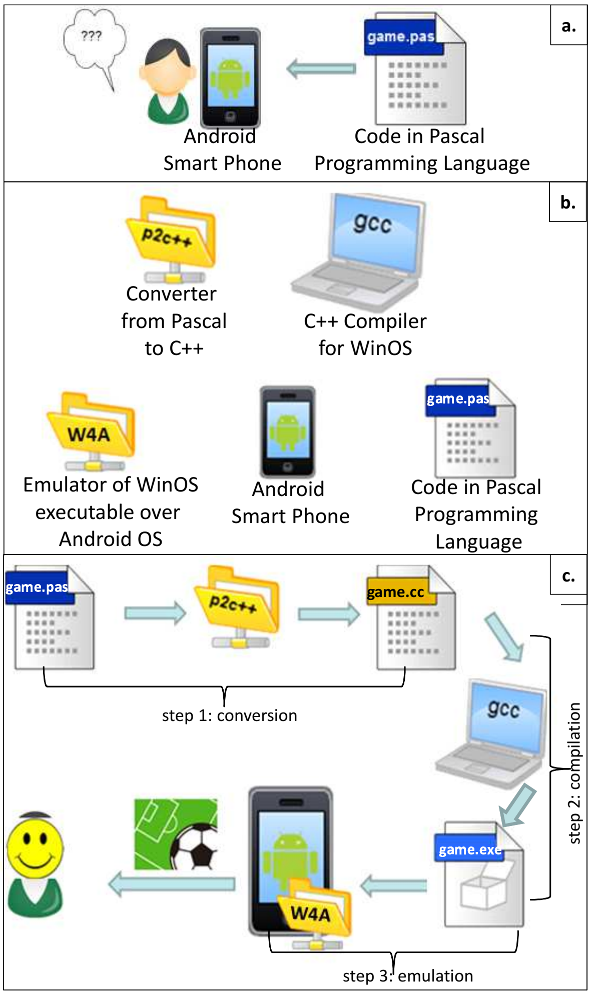


Figure 1: Running example. (a) The problem, (b) The available modules, (c) A series of conversions/emulations to achieve our objective

We can convey the main message of our approach through an example. Consider a user, say Yannis, who would like to compile and run on his mobile phone, software source code written many years ago, e.g. software code written in the Pascal programming language, stored in a file named `game.pas`. For example consider the situation illustrated in Figure 1a. *What can Yannis do? (to achieve his objective), What should we (as a community) do?, Do we have to develop a Pascal compiler for Android OS?, Do we have to standardize programming languages?* The direction and answer of the above questions (according to the approach that `Epimenides` follows), is that it is worth investigating whether it is already possible to compile and run that code on android by "combining" existing software, i.e. by applying a series of transformations and emulations. To continue this example, suppose that we have at our disposal only the modules that are shown in Figure 1b. Someone could then think that we could run `game.pas` on his mobile phone in three steps: by first converting the Pascal code to C++ code, then compiling the C++ code to produce executable code, and finally by running over the emulator the executable yielded by the compilation. Indeed, the series of conversions/emulations shown in Figure 1c could achieve our objective. However, one might argue that this is very complex for humans. Indeed this is true. We believe that such reasoning should be done by computers, not humans. `Epimenides` enables this kind of *automated reasoning.*

The contributions of this paper are:

- its presents `Epimenides`, a system offering novel interoperability reasoning services for digital preservation
- it presents an analysis of digital preservation scenarios of DANS, and shows how `Epimenides` could be used in these scenarios.

The rest of this paper is organized as follows. Section 2 discusses the context and the direction of this line of research. Section 3 presents the system `Epimenides`. Section 4 describes the scenarios provided by DANS and what `Epimenides` could do in each of them. Finally Section 5 concludes the paper.

2. CONTEXT, DIRECTION & RELATED WORK

The proposed methodology aims at offering a coherent approach for handling *interoperability dependencies.* Digital objects and digital collections should remain usable, i.e. one (human or artificial agent) should be able to understand and *use* the digital material over time. This is related to *interoperability*, and for this reason digital preservation has been termed *"interoperability with the future"*. Each interoperability objective or challenge (like those that were listed in [5], [9]) can be considered as a kind of demand for the *performability of a particular task* (or tasks). We can identify various tasks, which in many cases are layered. Examples of tasks include: *rendering* (for images), *compiling* and *running* (for software), *getting the provenance* and *context* (for datasets), etc. In every case the performance of each task has various *prerequisites* (e.g. operating system, tools, software libraries, parameters, representation information etc). We call these *dependencies.* The definition and adoption of standards (for data and services), aids interoperability because it is more probable to have (now and in the future) systems and tools that support these standards, than having systems and tools that support proprietary formats. From a dependency point of view, standardization essentially reduces the dependencies and makes them more easily resolvable; *even though it does not eliminate dependencies.* In all cases (standardization or not), we cannot achieve interoperability when the involved parties are not aware of the dependencies of the exchanged artifacts. However, the ultimate objective is the ability to perform a task, not the compliance to a standard, nor the availability of extra information. An important observation is that even if a digital object is not compliant to a standard, there may be tools and processes that enable the performance of a task on that object. However, as the scale and complexity of information assets and systems evolves towards overwhelming the capability of human archivists and curators (either system administrators, programmers or designers), it is important to aid this task, by offering services that can check whether it is feasible to perform a task over a digital object. For example a software written in 1986 could be executed on a 2013 platform, through a series of conversions and emulations. The process of checking whether this is feasible or not is too complex for a human and this is where *automated reasoning services* could contribute. Such services could greatly reduce the human effort required for periodically checking (monitoring) whether a task on a digital object is performable.

Towards this vision, in the context of APARSEN (Deliverable D25.2 [6]), past rule-based approaches for dependency management ([10], [8], [11]) were advanced for being able to capture converters and emulators. GapMgr[1] and PreScan[2] [7] are two systems that have been developed based on the dependency management model of past approaches [8], [11]. The new proposed modeling [6] enables the desired reasoning regarding task performability taking also into account the capabilities offered by *converters* and *emulators.* The prototype system `Epimenides` (which is the focus of the current paper) is the first system that realizes this approach and demonstrates its functionality.

Another related work is the TIMBUS[3] project. TIMBUS [2] is an EU co-funded project focuses on the preservation of business processes. It employs reasoning-based enterprise risk management to identify preservation risks, mitigation options and to determine the options' cost-benefit. It determines the metadata that needs to be captured and the dependencies (software and hardware components) of relevant process. However there are currently no publicly available TIMBUS software products that exploit this reasoning. In addition, there are several works that can assist various task of the digital preservation area. For example there are tools for the identification of file formats (e.g. DROID[4], Jhove[5], Apache Tika[6]), for getting the details about a technical environment (e.g. TOTEM [1], Preservation Network Model (PNM) [4]) and for getting assistance in preservation planning (e.g. Plato[3]). However none of the aforementioned works offers an automated reasoning for checking whether a task can be performed over a digital object, which is the ultimate objective in a digital preservation strategy.

3. THE EPIMENIDES PROTOTYPE SYSTEM

As stated `Epimenides` is the first system that realizes the approach described in [12, 6]. Its implementation is based on W3C standards (e.g. HTML, CSS, RDF, SPARQL), and its Knowledge Base (expressed in RDF/S) contains information about all MIME types and the modeling of various quite common tasks. Since it is based on Semantic Web technologies it can be straightforwardly enriched with in-

[1] http://athena.ics.forth.gr:9090/Applications/GapManager/

[2] http://www.ics.forth.gr/isl/PreScan

[3] http://timbusproject.net/

[4] http://digital-preservation.github.io/droid/

[5] http://jhove.sourceforge.net/

[6] http://tika.apache.org/index.html

formation coming from other external sources (i.e. other SPARQL endpoints).

Epimenides is a web accessible system[7], it can be used by several users (and each of them can define and maintain his/her own profile). Fundamental notions of Epimenides are *module*, *dependency* and *profile*. A module can be a software/hardware component or even a Knowledge Base (KB) expressed either formally or informally, explicitly or tacitly, that we want to preserve. A profile is the set of modules that are assumed to be known (available or intelligible) by a user, and this notion allows controlling the number of dependencies that have to be recorded formally.

3.1 Use Cases

In brief Epimenides offers the following services: (a) Task-Performability Checking, (b) Consequences of a Hypothetical Loss and (c) Identification of Missing Modules. A Use Case Diagram providing an overview of the supported use cases is given in Figure 2.

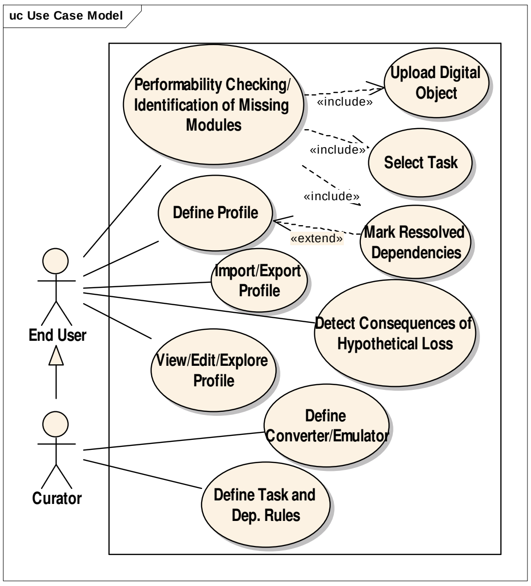

uc Use Case Model
Performability Checking/ Identification of Missing Modules
«include»
Upload Digital Object
«include»
Select Task
«include»
Define Profile
«extend»
Mark Ressolved Dependencies
Import/Export Profile
Detect Consequences of Hypothetical Loss
End User
View/Edit/Explore Profile
Define Converter/Emulator
Curator
Define Task and Dep. Rules

Figure 2: Use Case Diagram of Epimenides

3.2 User Interface

The user interface contains a menu divided in three sections as shown in Figure 3. The first section contains the option "Upload Digital Object" which is the core functionality of Epimenides. The "Manage Profile" section contains options for adding/deleting modules to/from a profile. Finally, the "Manage System" section contains options for curators that allow them to define Tasks, Emulators and Converters.

3.3 Performability Checking

To perform a task we have to perform other subtasks and to fulfil the associated requirements for carrying out these tasks. Epimenides is able to decide whether a task can be performed by examining all the necessary subtasks, exploiting also the possibilities offered by the availability of converters and emulators. In our example of Figure 1, the availability of a converter from Pascal to C++, a compiler of C++ over Windows OS and an emulator of Windows OS over Android OS, allows the inference that the particular Pascal file is runnable over Android OS.

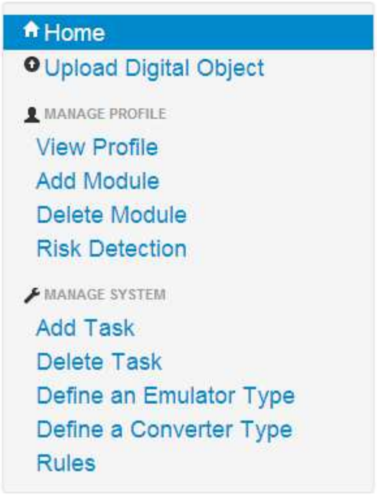


Figure 3: Main functionality of Epimenides

The core service of Epimenides, performability checking, is illustrated in the screenshots of Figure 4. After logging in to Epimenides, *the user can upload a digital object* (file or zipped files) and select the task whose performability he or she wants to check. The system checks the dependencies and computes the corresponding gap. To identify the dependencies of the uploaded objects, the system exploits the extension of the object (e.g. .pdf, .doc, .docx). An alternative way to identify file types that could be supported by Epimenides is to use *file format identification* tools like those that mentioned in Section 2. The KB of Epimenides contains the dependencies of some widely used file types. The identified dependencies are then shown to the user. The user can *add* those that (s)he already has, and this is the method for defining his/her profile *gradually*. In this way the user does not have to define a profile in one shot. The system stores the profiles of each user (those modules marked as "I have them") to the RDF triplestore.

3.4 Architecture and Current Deployment of Epimenides

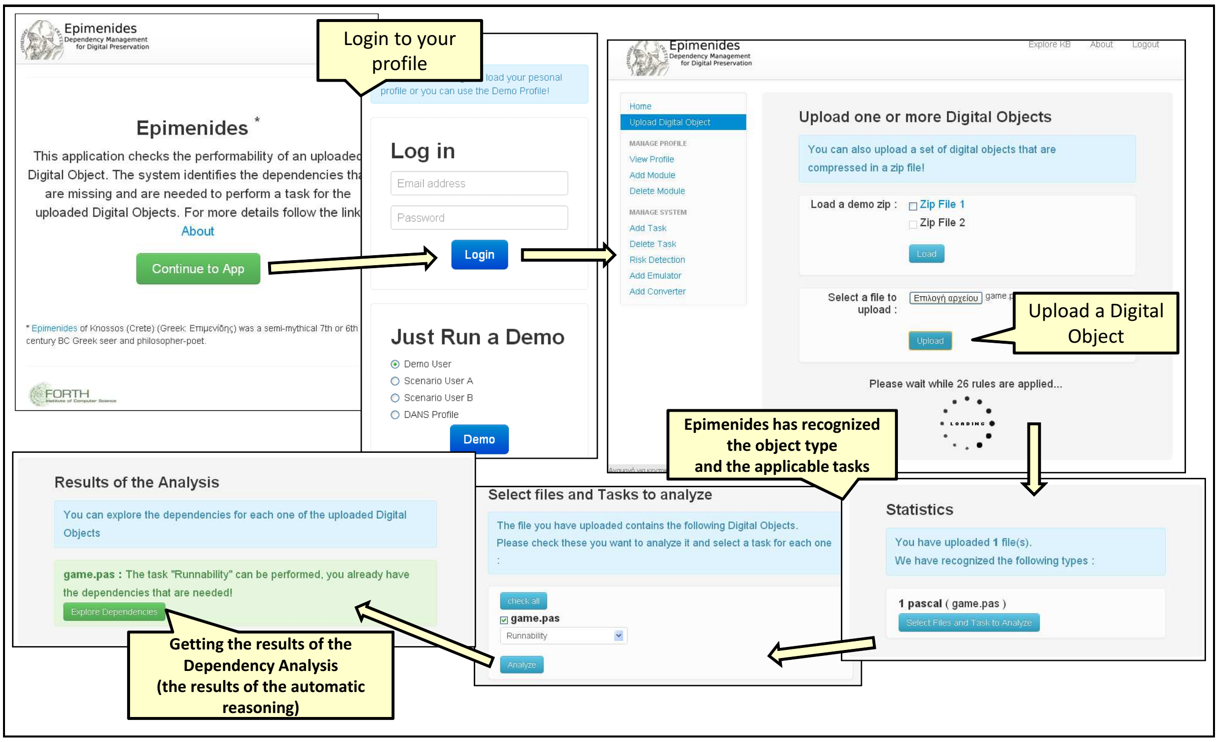
The server side of Epimenides is implemented in Java and it uses the Apache Tomcat[8] 7.0.3 web server. The used triple store is the OpenLink Virtuoso[9] 06.01.3127 version, and the Virtuoso Jena RDF Data Provider[10] is used for the communication with the triplestore. Figure 5 shows the component and deployment diagram of Epimenides. The architecture of Epimenides is based on the MVC (Model View Controller) pattern, meaning that all business logic is implemented in Java Servlets and all communication and data transfer issues are addressed with the use of Java Beans. The presentation of data is specified using JSP pages in order to separate the presentation design from the application logic, making easier the extension and modification of the system.

[7] http://www.ics.forth.gr/isl/epimenides/

[8] http://tomcat.apache.org/

[9] http://virtuoso.openlinksw.com/

[10] http://www.openlinksw.com/dataspace/doc/dav/wiki/Main /VirtJenaProvider



Epimenides Dependency Management for Digital Preservation

Epimenides *

This application checks the performability of an uploaded Digital Object. The system identifies the dependencies that are missing and are needed to perform a task for the uploaded Digital Objects. For more details follow the link About

Continue to App

* Epimenides of Knossos (Crete) (Greek: Επιμενίδης) was a semi-mythical 7th or 6th century BC Greek seer and philosopher-poet.

Login to your profile

load your pesonal profile or you can use the Demo Profile!

Log in

Email address

Password

Login

Just Run a Demo

● Demo User

○ Scenario User A

○ Scenario User B

○ DANS Profile

Demo

Explore KB About Logout

Home

Upload Digital Object

MANAGE PROFILE

View Profile

Add Module

Delete Module

MANAGE SYSTEM

Add Task

Delete Task

Risk Detection

Add Emulator

Add Converter

Upload one or more Digital Objects

You can also upload a set of digital objects that are compressed in a zip file!

Load a demo zip : □ Zip File 1 □ Zip File 2

Load

Select a file to upload : Επιλογή αρχείου game.p

Upload

Upload a Digital Object

Please wait while 26 rules are applied...

Epimenides has recognized the object type and the applicable tasks

Statistics

You have uploaded 1 file(s).

We have recognized the following types :

1 pascal (game.pas)

Select Files and Task to Analyze

Select files and Tasks to analyze

The file you have uploaded contains the following Digital Objects. Please check these you want to analyze it and select a task for each one :

check all

☑ game.pas

Runnability

Analyze

Results of the Analysis

You can explore the dependencies for each one of the uploaded Digital Objects

game.pas : The task "Runnability" can be performed, you already have the dependencies that are needed!

Explore Dependencies

Getting the results of the Dependency Analysis (the results of the automatic reasoning)

Figure 4: Checking the performability of a digital object

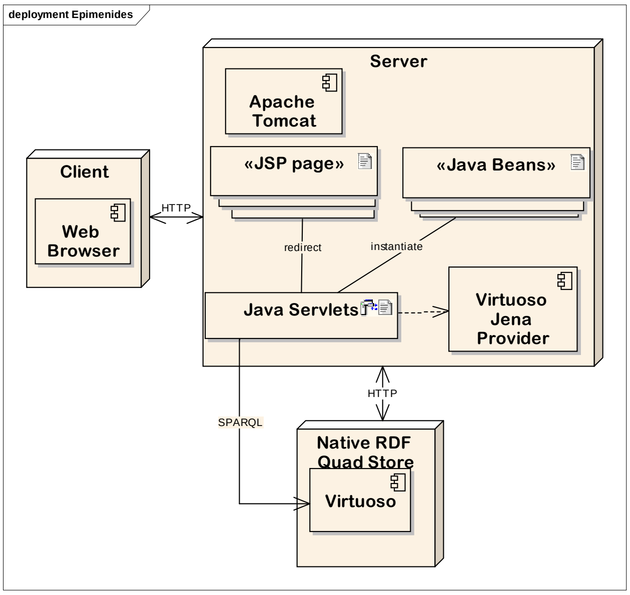


Figure 5: The deployment diagram of Epimenides

More information about the architecture of the Knowledge Base is given in [6].

4. EPIMENIDES USED BY A RESEARCH DATA ARCHIVE

We have conducted a case study in which the reasoning service of Epimenides is applied in the research data archive of DANS (Data Archiving and Networked Services, NL)[11].

DANS aims at promoting sustained access to digital research data. For this purpose, it encourages researchers to archive and reuse data in a sustained manner, e.g. through the online archiving system EASY[12]. DANS also provides access, via NARCIS[13], to scientific datasets, e-publications and other research information in the Netherlands. Apart from these, the institute provides training and advice, and performs research into sustained access to digital information.

Table 1 describes some of the common practices that are followed by curators of DANS in order to archive a file in the digital repository.

4.1 Scenarios

In collaboration with DANS, we have defined a number of scenarios that indicate where and how automatic reasoning related to long-term access to digital objects could be used. The analysis yielded five scenarios, whose description follows. In brief, the desired (for DANS) tasks are mainly related to the notion of *acceptable/preferred formats*, and with the runability of DANS software (including computability of checksums).

For each scenario there is a short *description* and an *applicability* subsection that discusses how the dependency management approach can be applied and how it can be realized by Epimenides.

4.1.1 Scenario 1: Supporting the notion of Preferred/ Acceptable Formats for Ingestion

[11] http://www.dans.knaw.nl/en

[12] http://easy.dans.knaw.nl

[13] http://www.narcis.nl

Table 1: Common practices that DANS follows in order to archive a file

Type of Data:	Common Practices
Documents	All documents (and also presentations - Powerpoint) are converted to PDF/A. For the conversion Adobe PDF convertor of Adobe Acrobat Professional is used.
Images/Illustrations	Both JPEG and TIFF (archival format) are used. Managing software: Adobe Photoshop.
Windows Metaformat (WMF) & Encapsulated Postscript (EPS)	Are converted by Adobe Illustrator to SVG (Scalable Vector Graphics) files.
Databases	dBASE (.dbf), Access (.mdb) and MS Excel Openoffice Calc are converted to CSV format. The export function of MS Access is used for the conversion. Some specific rules are applied (decimal delimiter, memo fields, double quotes in text fields). DBase (.dbf) files are imported in MS Access and exported in comma-separated values (.csv) files. Excel (.xls) files are exported to tab-delimited text files, then imported in MS Access and subsequently exported to comma-separated values (.csv) files.
Geographical Information files	Images such as Mapinfo Workspaces are converted to PDF/A. MapInfo TAB files are converted to MID and MIF files. ArcGIS Shapefiles are converted to MIF/MID by the Data Interoperability Extension of ArcGIS. Grid-fles are converted to ASCII-text files. MIG files are converted by the MAPINFO MIG-Toolbox. Surfer .grd and .srf files are converted by Golden Software Surfer to GS ASCII. Georeferenced images are converted by ArcMAP to a standard bitmap; this file is converted by Adobe Photoshop to JPEG and TIFF.
Computer Aided Design	AutoCAD files are stored as AutoCAD R12/LT2 DXF.

Description: For a number of data types (tables, text, images, etc.), specific file formats are considered to be durable at least into the near future. DANS maintains a list of *acceptable* and *preferred* formats. These lists are the basis for data archiving activities. The list that DANS currently uses is shown in Figure 6.

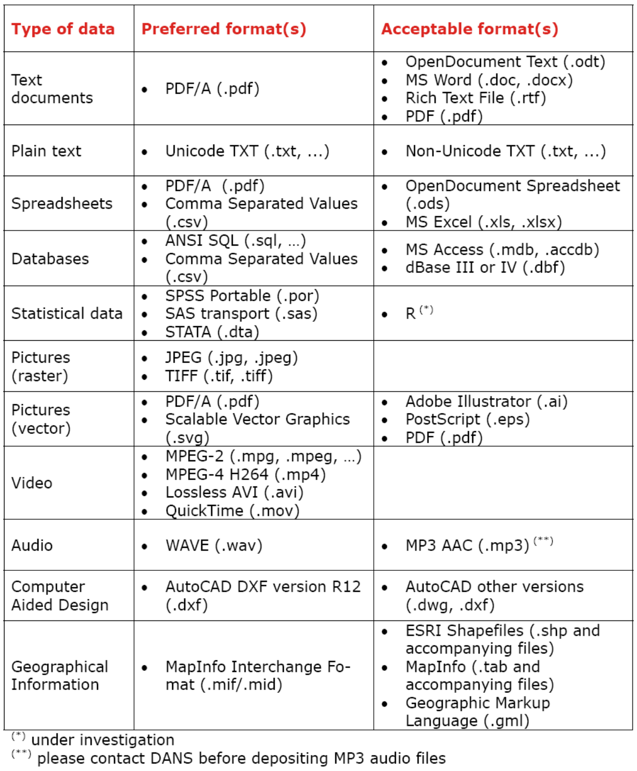
Applicability: If the converters (or emulators) that are in use by DANS for carrying out the migration activities, are registered in a system like `Epimenides`, then the system can be exploited not only for checking whether a newly ingested file is in an acceptable/preferred format, but also for checking whether it is migratable to one preferred or acceptable format using the migration/emulation software that DANS uses and has registered.

To realize this scenario, one has to define a profile (say $profile\_DANS$) that consists of:

i. The list containing the software that DANS uses for managing a file having an acceptable/preferred file format (e.g. `AcrobatReader` for rending PDF files, `VLC player` for playing mpg/mpeg/mp4/avi/mov files). At least one software tool per format is required.

ii. For each file type in the list of acceptable/preferred list, a task has to be associated (the one usually applicable to such file types) and the dependencies for that task have to be delivered in a way so that they are satisfied by the list of software described in (i). (e.g. for a pdf file type we can identify the *Rendering* task, and the need of (a) a pdf file, (b) an `AcrobatReader`).

iii. The list of tools that DANS uses for migration/conversion purposes (e.g. the tool `doc2pdf` for converting doc files to pdf).



Type of data	Preferred format(s)	Acceptable format(s)
Text documents	• PDF/A (.pdf)	• OpenDocument Text (.odt) • MS Word (.doc, .docx) • Rich Text File (.rtf) • PDF (.pdf)
Plain text	• Unicode TXT (.txt, …)	• Non-Unicode TXT (.txt, …)
Spreadsheets	• PDF/A (.pdf) • Comma Separated Values (.csv)	• OpenDocument Spreadsheet (.ods) • MS Excel (.xls, .xlsx)
Databases	• ANSI SQL (.sql, …) • Comma Separated Values (.csv)	• MS Access (.mdb, .accdb) • dBase III or IV (.dbf)
Statistical data	• SPSS Portable (.por) • SAS transport (.sas) • STATA (.dta)	• R [(*)]
Pictures (raster)	• JPEG (.jpg, .jpeg) • TIFF (.tif, .tiff)	
Pictures (vector)	• PDF/A (.pdf) • Scalable Vector Graphics (.svg)	• Adobe Illustrator (.ai) • PostScript (.eps) • PDF (.pdf)
Video	• MPEG-2 (.mpg, .mpeg, …) • MPEG-4 H264 (.mp4) • Lossless AVI (.avi) • QuickTime (.mov)	
Audio	• WAVE (.wav)	• MP3 AAC (.mp3) [(**)]
Computer Aided Design	• AutoCAD DXF version R12 (.dxf)	• AutoCAD other versions (.dwg, .dxf)
Geographical Information	• MapInfo Interchange Fomat (.mif/.mid)	• ESRI Shapefiles (.shp and accompanying files) • MapInfo (.tab and accompanying files) • Geographic Markup Language (.gml)

[(*)] under investigation
[(**)] please contact DANS before depositing MP3 audio files

Figure 6: DANS: Preferred and acceptable formats

Having completed these steps, the end user (or archivist) could use `Epimenides`. Whenever he uploads a file, `Epimenides` prompts the applicable task and directly informs the user if it is in an acceptable format or migratable to an acceptable format using the software that DANS has.

Without such facility it is difficult for a curator to (a) determine that an archived dataset is formatted in a durable format and (b) to have an overview of the applicable file format migration procedures that can be carried out to convert a file into a preferred file format (given that the list of preferred file formats will change over time as file formats become obsolete).

4.1.2 Scenario 2: Managing the set Preferred/ Acceptable File Formats

Description: As the usability and durability of file formats tend to change over time, for DANS it is important to periodically monitor and assess the applicability of the list of preferred formats and if it is necessary to replace a file format that became obsolete with a new one. Also new preferred formats can be introduced in the list. Specifically, say every year, the specifications on the list of preferred file formats have to be assessed based on a number of criteria (e.g. discussions in literature, consensus of organizations that provide guidelines in this field, etc.).

Applicability:

i. To add a new format in the list of acceptable/preferred file formats, the archivist can register it to the Knowledge Base of `Epimenides`. The check performed at ingestion will then function as expected (i.e. in accordance with the revised list of acceptable formats).

ii. Before deleting a file format (or managing software) from the list of acceptable/preferred file formats (or available software respectively), the archivist can check the impact of that deletion, i.e. the impact that this deletion will have on the performability of tasks over the archived files. This service (risk detection) is described in detail in [12].

iii. To remove a file format (or managing software) from the list of acceptable/preferred file formats (or available software respectively), the archivist can delete the corresponding entries from the system. After doing so, the checking at ingestion (Scenario 1) will function as expected, i.e. in accordance with the revised list of acceptable formats.

Without such services it is difficult to identify all the consequences of file format's obsolescence. It is also difficult to identify what will happen if managing software that is able to convert to/from a preferred file format, is lost or will become obsolete.

4.1.3 Scenario 3: Migration

Description: Research datasets are submitted in a number of formats to the data archive by the depositors. The data archive stores and manages these datasets in the format as submitted by executing the so-called "bit-preservation" (more about bit preservation in a next scenario). The data archive manages all formats but only commits itself to the long-term usability of files that are formatted according to the so-called preferred formats, described in the previous scenarios. In two situations a file format migration is required: (a) as part of the ingest procedure, files not formatted according to the preferred file format are migrated to a suitable preferred file format, (b) in case in the future a preferred file format becomes obsolete the files have to be migrated to this new format. The migration process requires using certain tools. Quality features of these tools are: speed, accuracy, level of completeness, and usability of the tool.

Applicability: The dependency management approach can show the archivist whether a file format migration is possible using the software that DANS has (recall Scenario 1). Also since a migration can be performed with different tools (or execution plans in general), the proposed system can assist the archivist by showing him/her, the possible actions/tools and this can be achieved by exploring the dependencies that the system offers.

Without this approach it is difficult for a human to identify all possible migration plans.

4.1.4 Scenario 4: Software Preservation

Description: Despite the fact that research data archives aim for durable access of datasets, there are cases where specific software is required to be able to use the datasets. For such cases, activities have to be undertaken to guarantee that this software is usable over time. Software preservation involves much more dependencies than research data preservation (e.g. changing operating systems, proprietary source code, etc.). Research data archives currently have no general accepted software preservation strategy.

Applicability: The example described in section 1 (Figure 1) falls in this scenario. Also [12] demonstrated this scenario with various examples.

4.1.5 Scenario 5: Authenticity of digital objects

Description: The bit preservation scenario involves activities to guarantee that digital objects do not become corrupted. This means that not one bit is changed over time. Thus the integrity of the data objects is guaranteed. This can be achieved by creating checksums on the occasion where the digital objects are ingested in the data archive and periodically checking whether a checksum is still valid. Dependencies in the scenario are the strength of the checksum procedures and the time interval the checksum is checked as part of the bit preservation activities.

Applicability: If checksums are supposed to ensure that the data have not been corrupted, an archive can model as task the computation of checksums for being sure that in the future the archiving organization will be able to recompute and compare them with the stored ones. Note that there are several tools for computing checksums[14]. We can say that this is a special case of Scenario 4.

4.2 Consolidation of the Scenarios

[14] `http://en.wikipedia.org/wiki/Checksum#Checksum_tools`.

Table 2: Application of the Methodology for the case of DANS

General Step	Specialization for the case of DANS
1. Identify the desired tasks and objectives	The desired tasks are: a. those related to the list of the acceptable/preferred formats, e.g. `render` (for pdf, txt, pictures), `play` (for video, aurio), `getTheRelationalModel` (for spreadsheets, databases), etc. b. those related to the runability of DANS software (including computability of checksums).
2. Model the identified tasks and their dependencies (check hierarchy)	Model the tasks using the list of software described in Scenario 1 (i). (Section 4.1.1). Moreover the dependencies of the runability of the tools that DANS uses for migration have to be modeled. Model the software dependencies that are required for running the software that DANS uses. In general the required modeling is quite simple, analogous to the examples given in [12].
3. Specialize the rule-based approach	It seems that there is not need for any particular specialization.
4. Identify Ways to capture dependencies (manual, auto, semi-automatic)	The file types are detected automatically (when one uses the upload feature of `Epimenides`). For applying this approach in big collections of files, various tools could be used for automating this process. Surely, in an operational setting the proposed functionality could extend or complement the functionality of the ingestion procedures of the systems that DANS currently uses.
5. Customize, use and exploit the dependency management services	For demonstration purposes this can be done using `Epimenides`, i.e. no need for customization or integration with the other systems of DANS. However, in an operational setting the processes and systems of DANS (EASY, NARCIS) should be considered.
6. Evaluate	This can be done using `Epimenides`.

Table 2 consolidates the key points of the above scenarios describing them based on the steps of a general methodology introduced in [12], for modeling, capturing and managing dependencies for the needs of digital preservation.

4.3 Defining the Profile of DANS in Epimenides

Following the implementation requirements of the scenarios that were described in Section 4.1, we defined a profile for the case of DANS. Specifically:

- We have registered (using `Epimenides`) the managing software that DANS uses in order to manage the preferred/acceptable files.
- We have identified and registered to the KB of `Epimenides` the tasks that make sense to apply in the list of the preferred/acceptable files.
- Finally the migration tools of DANS have also been registered to the DANS profile.

This profile is available in the registry of `Epimenides` and can be used by the archivists of DANS to exploit the benefits of the automatic reasoning approach that are described in the above scenarios. It defines 21 converters, 11 managing software tools, 4 tasks, and 44 rules. The representation of the profile as RDF triples is around 2,405 RDF triples. The numbers are summarized also in Table 3.

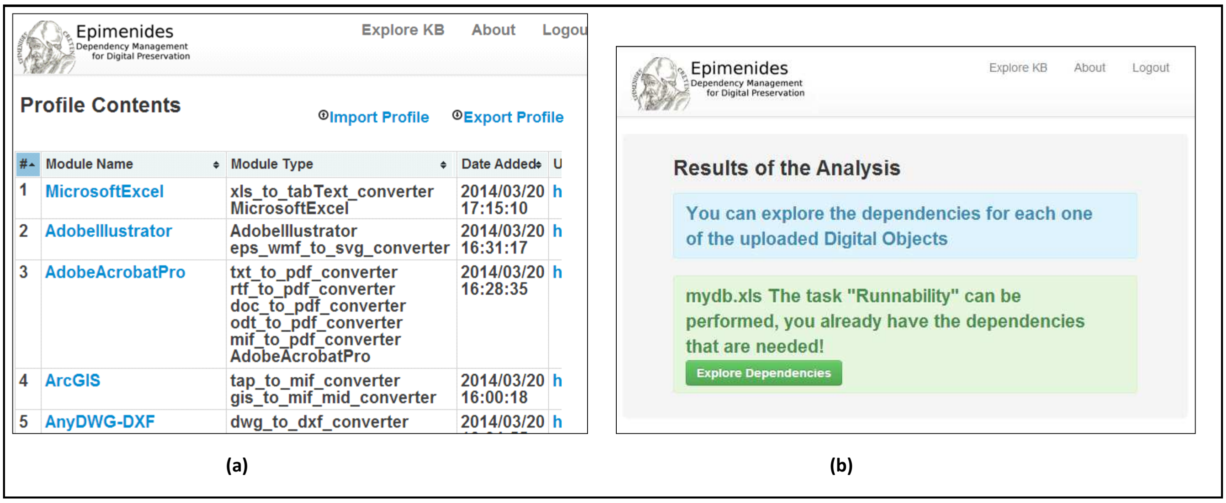
Table 3: DANS profile & Numbers

Component:	#
Converters	21
Managing Software	11
Tasks	4
Defined Rules	44
Triples in Repository	2,405

Considering the practices shown in Table 1, note that `Epimenides` with the DANS profile behaves as expected. For example, the practices of DANS for excel database files as described in Table 1 are: *"Excel (.xls) files are exported to tab-delimited text files, then imported in MS Access and subsequently exported to comma-separated values (.csv) files"*. Now suppose that we want to check if DANS could manage a database excel file, say `mydb.xls`. Two conversions should be applied according to the practice that is described before (.xls $\xrightarrow{MSExcel}$ .tab $\xrightarrow{MSAccess}$ .csv). Having defined (as shown in Figure 7a) in `Epimenides` that DANS holds the needed converters (`MS Excel` and `MS Access`) and uploading the `mydb.xls` to the system we can see in Figure 7b that the proposed automated reasoning has been applied and the appropriate tasks can be performed for this file.

5. CONCLUDING REMARKS

Digital material has to be preserved not only against loss or corruption, but also against changes in its ecosystem. In this paper we described `Epimenides`, a system that realizes an automatic reasoning approach for assisting this digital preservation problem. The approach is based on the description of *dependencies* that are required in order to achieve a *task*. `Epimenides` can be used by digital archives and digital libraries to help archivists in checking whether the archived digital artifacts remain *intelligible* and *functional*, and in identifying the consequences of probable losses.



Profile Contents

#	Module Name	Module Type	Date Added	U
1	MicrosoftExcel	xls_to_tabText_converter MicrosoftExcel	2014/03/20 17:15:10	h
2	AdobeIllustrator	AdobeIllustrator eps_wmf_to_svg_converter	2014/03/20 16:31:17	h
3	AdobeAcrobatPro	txt_to_pdf_converter rtf_to_pdf_converter doc_to_pdf_converter odt_to_pdf_converter mif_to_pdf_converter AdobeAcrobatPro	2014/03/20 16:28:35	h
4	ArcGIS	tap_to_mif_converter gis_to_mif_mid_converter	2014/03/20 16:00:18	h
5	AnyDWG-DXF	dwg_to_dxf_converter	2014/03/20	h

(a)

Results of the Analysis

You can explore the dependencies for each one of the uploaded Digital Objects

mydb.xls The task "Runnability" can be performed, you already have the dependencies that are needed!

(b)

Figure 7: a)Contents of DANS profile as shown in `Epimenides` b)Checking the performability of an excel file in DANS profile

In this paper we described (in the form of scenarios) how the reasoning service of `Epimenides` can be applied in the DANS data archive. We showed how various real activities are actually dependency management activities. Finally for the realization of the scenarios, we defined in `Epimenides` a profile for DANS.

From the technical side, an objective for future research is to develop quality-aware reasoning for enabling quality-aware preservation planning.

Acknowledgements

Work done in the context of NoE APARSEN (Alliance Permanent Access to the Records of Science in Europe, FP7, Proj. No 269977).

6. REFERENCES

[1] David Anderson, Janet Delve, L Konstantelos, A Ciuffreda, and M Dobreva. Totem: Trusted online technical environment metadata: a long-term solution for a relational database/rdf ontologies. 2011.

[2] José Barateiro, Daniel Draws, Martin Alexander Neuman, and Stephan Strodl. Digital preservation challenges on software life cycle. In *Software Maintenance and Reengineering (CSMR), 2012 16th European Conference on*, pages 487–490. IEEE, 2012.

[3] Christoph Becker, Hannes Kulovits, Andreas Rauber, and Hans Hofman. Plato: a service oriented decision support system for preservation planning. In *Proceedings of the 8th ACM/IEEE-CS joint conference on Digital libraries*, pages 367–370. ACM, 2008.

[4] Esther Conway, Matthew Dunckley, Brian McIlwrath, and David Giaretta. Preservation network models: Creating stable networks of information to ensure the long term use of scientific data. *Proc. PV2009, Madrid, Spain*, pages 1–3, 2009.

[5] Alliance for Permanent Access to the Records of Science Network (APARSEN). “D25.1 Interoperability Objectives and Approaches”, 2013. (http://www.alliancepermanentaccess.org/wp-content/uploads/downloads/2013/03/APARSEN-REP-D25_1-01-1_7.pdf).

[6] Alliance for Permanent Access to the Records of Science Network (APARSEN). “D25.2 Interoperability Strategies”, 2013. (http://www.alliancepermanentaccess.org/wp-content/uploads/2013/10/APARSEN-REP-D25_2-01-1_7.pdf).

[7] Y. Marketakis, M. Tzanakis, and Y. Tzitzikas. PreScan: Towards Automating the Preservation of Digital Objects. In *Procs of the International Conference on Management of Emergent Digital Ecosystems MEDES’2009*, Lyon, France, October, 2009.

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

[9] Y. Tzitzikas and B. Bazzanella. Interoperability Objectives and Approaches:Results from the APARSEN NoE . In *Proceedings of the 10th Annual International Conference on Digital Preservation (iPres2013)*, 2013.

[10] Y. Tzitzikas and G. Flouris. “Mind the (Intelligibily) Gap”. In *Procs of the 11th European Conference on Research and Advanced Technology for Digital Libraries, ECDL’07*, Budapest, Hungary, September 2007. Springer-Verlag.

[11] Y. Tzitzikas, Y. Marketakis, and G. Antoniou. Task-based Dependency Management for the Preservation of Digital Objects using Rules. In *Procs of 6th Hellenic Conf. on Artificial Intelligence, SETN-2010*, Athens, Greece, 2010.

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

A Persistent Identifier e-Infrastructure

Barbara Bazzanella
University of Trento
Via Sommarive 5, I-38123
Trento, Italy
barbara.bazzanella@unitn.it

ABSTRACT

Persistent identifiers (PIDs) have been recognized as a crucial enabling component for 2020 e-science infrastructures[1], having the potential of providing global keys for information access, reuse and exchange and creating a complex network of links which connect all the relevant entities in the research data landscape (e.g. digital objects to authors and datasets, authors to institutions and projects, projects to research products and fundings). The creation and full exploitation of this valuable network of connections is currently hindered by the fragmentation and lack of coordination of the persistent identifier ecosystem. Several initiatives have emerged with the aim of offering global identifier repositories for digital and non-digital entities but they are still focused on the needs of specific communities and the lack of interoperability between them is one of the major hurdles for the development of a globally connected scholarly infrastructure. The aim of this paper is to propose a Persistent Identifier e-infrastructure (based on an identifier service called Entity Name System) which provides a technical layer of interoperability which allows current identifier systems to interoperate and be coordinated across geographical, temporal, disciplinary, organization and technological boundaries. The Persistent Identifier interoperability e-infrastructure is presented as a cross-cutting core service enabling the development of advanced added-value services tailored to the specific needs of different communities and stakeholders of the e-science environment.

General Terms

Infrastructure

Keywords

persistent identifier e-infrastructure, interoperability, e-science research infrastructures, Entity Name System

1. INTRODUCTION

Science is global in scope, but it is only recently with the development of advanced information and communication technologies, that science is becoming global in practice. ICT-based infrastructures for science (i.e. e-science infrastructures) are at the root of this process, promoting the realization of an integrated information space where researchers can cooperate and share resources independently from their geographical location, and the access to increasing volumes of data and their processing is facilitated and empowered, making science more efficient and innovative. These infrastructures provide tools and services to support the full life cycle of scientific data (to gather, capture, transfer and process data), the dissemination of data across the boundaries of nations and scientific disciplines, the cross-linking of data in the digital space, the integration between scientific data and publications. According to the framework proposed by the High Level Expert Group of Scientific data [11], e-science infrastructures can be seen layered systems where different actors, data types and services interrelate within a global space and community services specific to each community or discipline rest upon common low level services cutting across the global system. A solid infrastructure for managing unique identifiers for all the entities involved within the global scientific data infrastructure - including digital objects, authors, contributors, datasets, funding agencies, projects and many others - is a a critical low level service to provide the layer of interoperation and trust of data necessary to enable access, use, reuse and exchange of data (see Figure 1) in a collaborative integrated research environment [5].

However, since a number of different identifier systems with different scope, level of maturity and technical sophistication are already in use by different communities and no single integrating identifier system seems meet the needs of all the communities and provide a service to identify all the relevant entities which populate the articulated network of connections within the research arena, the identifier infrastructure should not only provide a layer for assigning identifiers to resources and managing them, but it should provide an interoperability infrastructure which makes existing identifier systems able to interoperate and be integrated without the need to introduce a further identification solution in addition to those already consolidated and adopted by the different communities. The development of an interoperable identifier infrastructure is an essential step for unlocking the value of research data and creating a digital globally connected

[1] http://ec.europa.eu/programmes/horizon2020/en/h2020-section/european-research-infrastructures-including-e-infrastructures

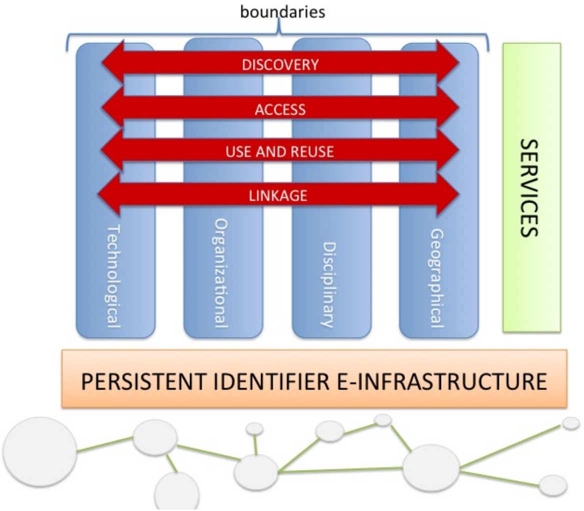


Figure 1: A Persistent Identifier e-Infrastructure

research environment in the near future. Even though, as pointed out in the DIGOIDUNA study [5], this is far from being a merely technical issue, opening a multidimensional spectrum of challenges dealing with economic, societal and policy aspects which need to be integrated into a coordinated model, the technical implementation of the agreed framework is an unavoidable step to secure the concrete and efficient operation of the infrastructure. This paper proposes a technical infrastructure exploiting an existing solution for managing global identifiers (called Entity Name System) which aims to provide a technical layer of interoperability allowing current identifier systems to interoperate and be coordinated across geographical, temporal, disciplinary, organization and technological boundaries. The Persistent Identifier interoperability e-infrastructure is designed as a cross-cutting core service enabling the development of advanced added-value services tailored to the specific needs of different communities and stakeholders of the e-science environment.

2. FROM URLS TO PERSISTENT IDENTIFIERS

The ability to reliably identify and locate digital information over time has become increasingly relevant in recent years in distributed digital environments. The Web infrastructure offers a very direct way to locate digital information based on the Uniform Resource Locator (URL). The URL specifies the physical location on a particular server from which to retrieve the digital resource (which could be a digital document, a dataset, an image, a video or any other digital resource on the Internet). However, since the Web is highly dynamic and resources are often moved to different locations during their lifecycle, the identification of digital content through URLs has proven to be a very fragile mechanism. When a digital object is transferred to a different destination or it goes off-line, the corresponding URL ceases to identify and locate the object and the link becomes “a broken link”. Moreover, if the location where the object was initially stored, is subsequently occupied by a different object, the corresponding URL could be used to locate two different resources at two different moments of time. This explains why URLs are only temporary identifiers and cannot be used to provide ongoing access to digital resources.

Persistent Identifiers (PIDs) have been introduced as a solution to address this issue providing an identification mechanism in which the identifier is not strictly bound to a specific digital location. Unlike a URL, a persistent identifier is a permanent association between a unique name and an information object which can be the resource itself or a representation of it (i.e. metadata describing it). This association is maintained independently of the physical location of the information object. If the location changes, the persistent identifier still remains the same providing a different way to retrieve the resource (e.g. a different URL where the object is placed) or an appropriate representation of the resource. Indeed persistent identifiers can be used to identify both digital and non digital entities (e.g. people). Even though at first persistent identifiers were mainly used for identifying digital content (publications and scholarly works for example), it has become increasingly evident that many non-digital resources need to be uniquely identified in order to extract value from the representation of digital assets. In the scholarly domain, for example, the need to unambiguously represent authors and contributors and associate them with their scientific outputs (e.g. publications, datasets, software), has favored the development of several author identification systems. More recently, other initiatives like the I^2 (Institutional Identifiers) working group[2] have started to define a standard for an institutional identifier by proposing to leverage existing solutions like ISNI.

Many different persistent identifier solutions (e.g. URN, Handle, DOI, ARK, PURL, ISNI, ORCID) have been proposed in recent years which aim to reproduce in the digital environment the two main functions that traditional identifier systems provide in other cultural contexts (like identifiers for books in traditional libraries), i.e. **identification** and **access**. Identification means using a label to name an object and distinguish it from other similar objects. Persistent identifiers aim to identify resources in 1) unique, 2) location-independent, 3) persistent way. This means that 1) a persistent identifier is only assigned to a single object and never reused within the domain of creation, 2) a persistent identifier is not intrinsically bound to the location of the object; 3) the association between the identifier and the object should be maintained over time. Identifiers that are designed simply to identify resources have little utility in the digital world. The second requirement of persistent identifiers is that they operate as durable keys to access to digital content. As we have stated above, access to the identified resources (or information about them) should be guaranteed over time. This is usually realized through different strategies, like a layer of indirection within the HTTP protocol (e.g. PURL, ARK), a resolver mechanism dissociated from the HTTP protocol (e.g. Handle, DOI, URN) or conferring stability to Web identifiers (e.g. Cool URIs). More importantly persistent access is ensured thorough a complex social and organizational infrastructure of policies and rules involving registration agencies and content providers (see for example the social infrastructure of registration agencies coordinated by the International DOI Foundation which reg-

[2] `http://www.niso.org/workrooms/i2`

ulates the DOI system).

2.1 The current landscape of Persistent Identifiers in science

Identification and long-term accessibility are fundamental in most sectors of human activity, but are crucial for scientific information management especially in recent years due to the rising growth of scientific production, the digitization of content and the distribution of data and services across different systems and networked infrastructures.

The consistent adoption and use of persistent identifiers is a critical step for all the main phases of scientific production and fruition of its products on a global scale. Experimental data should be collected, discovered and shared within a global scientific community and across different science domains, data should be uniquely attributed to the people who contributed to their generation and connected with scientific works, projects and publications. Authors should be uniquely identified across disciplines and other boundaries and associated with their entire scientific production and linked to their professional activities (e.g. projects, events, teaching experiences) and membership institutions. Persistent identifiers have been recognized as fundamental building blocks for enabling accessibility, trustworthiness, provenance and quality assessment in e-science. This explains why assessing the impact of the use of different identifier solutions for digital objects, authors and other relevant entities has become a critical issue for policy makers and funding agencies especially when they aim for the realization of large-scale ICT infrastructures for e-science as the fundamental scientific production environment. This attention is confirmed by the recent EU Framework Program for Research and Innovation (Horizon 2020) in the area of Research Infrastructures[3], which envisions the development of a digital identifier infrastructure for digital objects and authors as a core service across e-infrastructures.

However, widespread adoption of persistent identifiers is far from being realized and the level of maturity and technical sophistication of the current identification solutions is widely diversified. While identification systems are well established in some specific domains and for certain kinds of resources (e.g. DOI for scholarly and scientific publications, URN for digital resources in many libraries and institutional repositories, ARK for digital objects in traditional and digital libraries), persistent identifiers are only recently (and quite slowly) emerging for other entities in the scientific domain. The introduction of non-ambiguous and persistent identifiers for authors and contributors is quite a recent practice, which have started to produce a number of local (sometimes national) ad hoc solutions in specific domains or systems (e.g. DAI in the Dutch Research System, author identifiers in arXiv, Scopus Author id developed by Elsevier, ResearcherID developed by Thomson Reuters). It's only recently that we are assisting to the development of more global integrating solutions for identifying authors and contributors across systems (e.g. ISNI, ORCID). Other identifier solutions (e.g. DOI through DataCite) have started to be adopted for identifying complex scientific entities, like datasets. Even more recent are persistent identifiers for institutions (e.g. Ringgold in the publisher domain). Another aspect of the current persistent identifier solutions is that resources can be part of different domains and can be identified by different identifiers in different systems. The same digital object which is assigned a DOI in the publishing domain can be assigned a URN within an institutional repository. Nowadays there is no overall integrating solution to map and retrieve different identifiers for the same resource and link a resource to all the entities (in turn identified by other persistent identifiers) with which it is interconnected. This makes hard to reuse identifiers across domains, integrate metadata from different sources and create integrating cross-boundary services based on different identification systems.

From this brief overview, two aspects of the persistent identifier landscape in e-science emerge: 1) **the fragmentation of the ecosystem** populated by a number of identifier solutions not equally diffused and consolidated 2) **a lack of an interoperability solution** for current persistent identifier systems which are nowadays difficult to integrate to offer interconnected services.

2.2 Toward Interoperability for Persistent Identifiers

In the last few years a number of initiatives and projects have started to create the ground for the realization of a global interoperable e-science framework based on the interoperability between identification systems. A study conducted on behalf of the European Commission, named DIGOIDUNA [5], has investigated the fundamental role of digital identifiers as enablers of value in e-science infrastructures and has performed a detailed analysis of strengths, weaknesses, opportunities and threats of the current digital identifier landscape in order to identify the main challenges and a set of recommendations which policy makers and relevant stakeholders should address to develop an open and sustainable persistent identifier infrastructure supporting information access and preservation. One of the main conclusions of the study is that to transform digital identifiers from simple means to manage data to keys for supplying knowledge and deliver value to the stakeholders within the research production, it is necessary to foster the development of an interoperable, cross-domain infrastructure for persistent identifiers supporting data access and sharing across national, organizational, disciplinary and technological boundaries. The implementation of this infrastructure poses several technical challenges but raises also a multidimensional spectrum of organizational, social and economical issues which should be addressed to ensure a coordinated ecosystem. Within the APARSEN project, the research on persistent identifiers has focused mainly on the definition of an interoperability framework for persistent identifier systems [1] which defines some key assumptions and requirements to identify the trustable candidate systems which can take part in the framework, an ontology which specifies the structure of data and the core set of relationships linking the identified entities within the framework and finally a small set of services which can be implemented on top of the framework. A demonstrator has also been developed to provide evidence of the potential applicability of the model and related basic services [2].

[3] `http://ec.europa.eu/research/participants/data/ref/h2020/wp/2014_2015/main/h2020-wp1415-infrastructures_en.pdf`

Other initiatives have started to define cooperation agreements and complementary architectures to ensure interoperability between independent systems or organizations. ORCID and ISNI for example have made a first advance in this direction by rendering ORCID compatible with the ISNI ISO standard and assigning a block of numbers for identifying ORCID entities which cannot be reassigned by ISNI to different people[4]. The integration between Researcher ID and ORCID is another example of a bi-directional integrating initiative aimed at making information on the two systems interoperable and complementing. Similarly, the ODIN project[5] aims to define a roadmap for the integration and scalability of the DataCite and ORCID identifiers solutions to create a layer of interoperability between persistent identifiers for researchers, research works and their outputs (publications and data) in order to address four main challenges concerning research data management: accessibility, discovery, interoperability and scalability. The proposed solution is based on a conceptual model of interoperability [3] for linking research data and their contributors (embedding the corresponding PIs into metadata) through the coordination and alignment of the information flow across data centers, DataCite, and ORCID. The RDA PID Interest Group[6]is another example of the recent effort of coordinating the use of persistent identifiers for supporting referencing and citation of research products and their authors and contributors and manage the lifecycle of research data production.

Finally, other initiatives have been started within specific communities. In the library domain, the BIBFRAME initiative[7] has defined a lightweight framework (metamodel) for bibliographic description based on linked data principles to improve the integration, discoverability and reuse of library resources and their descriptions in a networked distributed environment. At the core of the proposed data model, there is the concept of BIBFRAME authority which is a resource representing a person, organization, place, topic, temporal expression and other entities associated with a BIBFRAME Work, Instance, or Annotation (i.e. the remaining classes of the model). BIBFRAME authorities are used not only to identify (via URIs) the above mentioned entities within the description, but also to link to external resources (for example traditional authorities) referring to the same entities by including their corresponding IDs. In this way, the mechanism of BIBFRAME authorities should provide a common lightweight interoperability layer over different Web-based authority resources connceting a BIBFRAME resource, such as a Work or Instance, and one or more authorities for related entities, such as a person, organization, or place, identified by other identifiers systems like a ID.LOC.GOV, ISNI, VIAF and others.

All these initiatives have the merit of having increased the awareness and consensus among relevant stakeholders and communities about the crucial role of a coordinated ecosystem of persistent identifiers at the heart of a global infrastructure for e-science. A lot of work has been done to define common objectives and share conceptual models and strategies to solve the persistent identifier interoperability problem. However, a solid technological solution for interoperating identifiers for digital objects, contributors, authors and other relevant entities is still lacking in the effort to develop a sustainable infrastructure providing a core layer of interoperability on which cross-cutting advances services for science and education can be implemented to encourage openness and collaboration across disciplines, communities and geographical boundaries. Based on the valuable results of the above mentioned initiatives, but also exploiting the experience on persistent global identifiers gained in the course of the OKKAM FP7 project[8], this paper addresses the same problem from a slightly different perspective, proposing a technical solution to implement a persistent identifier interoperability core service for e-science infrastructures. In the next section we start to describe the three main functionalities which should be supported by this core service.

3. INTEROPERABLE PERSISTENT IDENTIFIERS AS VALUE ENABLERS OF E-SCIENCE INFRASTRUCTURES

Interoperable persistent identifiers are key building blocks in managing the complex information space of e-science infrastructures and extracting value from it. We have identified three main core functionalities which explain this crucial role.

1. Ensuring and enhancing the persistent access, use and reuse of resources or related information across different boundaries (e.g. technological, disciplinary, institutional).

2. Providing the means for explicitly representing the network of relationships among all the relevant entities in the research landscape (authors, contributors, publications, data, research projects, grants, intitutions) and creating an integrated information space which can be walked through starting from any of the links and from which new knowledge can be formed.

3. Enabling the development of added-value services on top of integrated digital information spaces.

The maintenance of a solid relationship between the identifier and the associated entity, digital (e.g. an electronic publication) or non-digital (e.g. the author of the publication) is the fundamental mechanism to ensure persistent access and reuse of the resource itself or information related to it. This stable association is what confers persistence to the identifier. In an interoperability infrastructure this means not only guaranteeing the persistent link between a given identifier and the identified resource, but also managing possible alternative links (implemented by other identifier systems) which may provide a continued alternative access to the resource in case the first connection is not accessible (e.g. broken link or denied access permission). This means that the infrastructure should be able to connect identifiers for the same entity across different systems. Such a requirement can

[4] http://orcid.org/blog/2013/04/22/orcid-and-isni-issue-joint-statement-interoperation-april-2013

[5] http://odin-project.eu/

[6] https://rd-alliance.org/internal-groups/pid-interest-group.html

[7] http://www.loc.gov/bibframe/

[8] http://project.okkam.org/

be addressed, for example, by managing matching functionalities with allow to identify "same-as" relationships between persistent identifiers, i.e. two identifiers refer to the same entity. For example, given a DOI for an article the identifier interoperability infrastructure could provide access to the identified publication through a redirection mechanism which involves the DOI resolver, but could also provide alternative persistent identifiers for the resource, if any, (for example an URN or an ARK), giving alternative ways to access the target information object.

The implementation of this coreference mechanism has been largely discussed within WP22 of the APARSEN project and has been included as one of the fundamentals of the framework. In the APARSEN framework, coreferences between persistent identifiers (and the identity between the referents) are not inferred based on matching on metadata information describing the identified entities, but are directly extracted from the information object. Since often resources are identified by more than one PID (e.g. a document can be identified by a DOI and by a URN) and the presence of alternative identifiers can be made explicit in the metadata provided by the persistent identifier management systems (e.g. in the DOI kernel metadata the "referentIdentifiers" element is used for this purpose), the framework, and the related demonstrator, rest on the idea that the co-existence of two or more identifiers in the metadata about the entity can be exploited to automatically generate trusted identity relationships between information objects, by transitivity.[9] In brief, these are the only trusted co-references according to the APARSEN approach and they can be reliably used to integrate information across PID domains. This cautious approach has the advantage to reduce the risk of generating false positive matches, due to the fact that the matching process is based on the coreference information directly provided by trusted PID domains, but has the disadvantage to exclude from the integration process all the objects not linked through the inferred coreference chains. Since, as we have stated above, the use of PID is largely fragmented and inadequate for many entities potentially relevant for the e-science domain, it is difficult to imagine a broad applicability of the proposed approach to include the entire spectrum of entity types of interest.

In order to exploit the value of e-infrastructure data, it is necessary to have stable access not only to the single resources but also the relationships among these resources [10], like an author and his/her research output or the publications related to a given dataset. According to this perspective, a second element of value of managing persistent identifiers deals with making explicit and reusable the relations between the relevant entities within the scientific data infrastructure[9]. Again this can be realized making interoperable identifier systems for different types of resources, like those for authors and contributors with those for digital objects. The persistent identifier interoperability infrastructure should be able to provide the identification capabilities necessary to represent structured knowledge that can be integrated across systems and used to discover new elements of knowledge by querying and navigating the information space. For example, data providers should be able to represent their data and metadata by reusing identifiers already assigned to the relevant entities instead of assigning new identifiers. A dataset should be not only identified by a unique ID but should also be related to its author as part of its metadata. If the author has already been assigned an author ID registered within the infrastructure identifier registry, it is crucial that the data center can reuse the same ID for uniquely identifying the dataset since through it many relevant relationships can be inferred (for example that among the author publications there is one article based on the experimental results on the dataset).

The interoperability infrastructure for persistent identifiers is also crucial for the development of community added-value services which can be build on top of the (now fully) accessible scientific data and network of relationships around them. Due to the interoperability layer not only the information is extracted and integrated across systems but also the higher level services based on this information can interoperate and produce additional value, for example by facilitating the sharing of research findings, improving accessibility to research products and identifying authors and contributors of scientific outputs. For instance, enabling automatic discoverable connections between relevant entities participating in the scientific production value chain, like funding agencies, grants, projects, contributors, institutions and many others, research administration services for assessing the impact of research programs can be developed and provide a valuable instrument for research funders and policy makers.

From this perspective, identifiers and metadata enriched by uniquely identified information are value enablers of e-science infrastructures, by increasing the interoperability of data, facilitating the access to relevant and trustable information, increasing the trustworthiness of sources, revealing links and dependencies between data and solving ambiguity issues.

4. THE ENTITY NAME SYSTEM

The aim of this paper is to propose a technical solution to implement the layer of interoperability for persistent identifiers in e-science infrastructures. This solution is based on the Entity Name System (ENS) prototype developed in the context of the EU-funded project OKKAM[10]. The ENS[11] is a scalable infrastructure for assigning and managing unique identifiers for entities in decentralized distributed information environments like the Web and foster their global reuse. The first prototype of this system has emerged as a solution to the entity identification problem in the Semantic Web [6] and in other distributed contexts, that is the problem of integrating information about entities which are assigned different identifiers in different systems or by different users [7]. In order to deal with this problem, the ENS provides a

[9] Assuming for example that an object, say o1, is identified by a DOI and another object, say o2, is identified by the same DOI as o1 and by an ARK, the ARK of o2 can be used to derive the identity relation between o1 and a third object, o3, identified by the same ARK as o2, by transitivity of the identity relation. In this way chains of coreferences can be automatically generated (provided that the metadata information from different PID domains is structured in a common way) by simply trusting the coreferences included in the information objects.

[10] http://project.okkam.org/

[11] http://api.okkam.org/

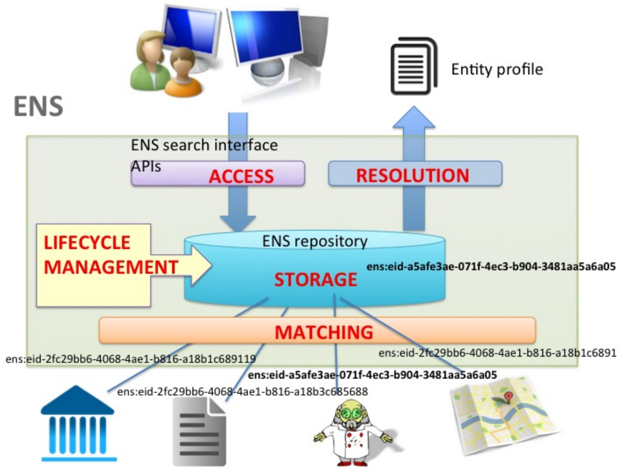

Entity profile
ENS
ENS search interface
APIs
ACCESS
RESOLUTION
LIFECYCLE MANAGEMENT
ENS repository
STORAGE
ens:eid-a5afe3ae-071f-4ec3-b904-3481aa5a6a05
MATCHING
ens:eid-2fc29bb6-4068-4ae1-b816-a18b1c689119
ens:eid-2fc29bb6-4068-4ae1-b816-a18b1c6891
ens:eid-a5afe3ae-071f-4ec3-b904-3481aa5a6a05
ens:eid-2fc29bb6-4068-4ae1-b816-a18b3c685688

Figure 2: ENS Infrastructure

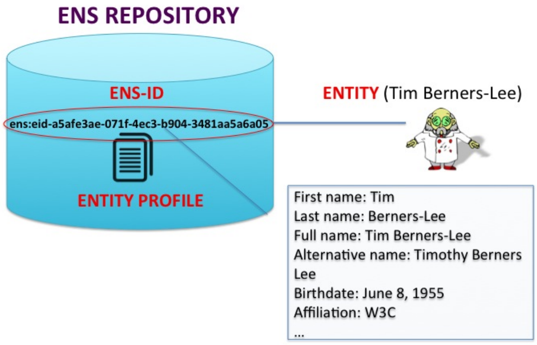


Figure 3: ENS Repository

service to assign global unique identifiers to entities named in information sources and reuse these identifiers across systems boundaries regardless of the place or domain where they have been first assigned. To this purpose the ENS has a repository for storing entity identifiers along with a short set of descriptive metadata, i.e. an entity profile, which is used with the aim to disambiguate each entity from the others. When a human user or an application searches the system for an identifier (for example by keywords), information in the entity profiles is used to establish (through advanced entity matching algorithms) if an identifier has been assigned and stored for that entity. Otherwise, a new identifier is minted and returned by the system. The systematic reuse of the identifiers created and maintained in the ENS would reduce the multiplication of identifiers for entities and enable a frictionless entity-centric integration of information spread and scattered on the Web. The ENS infrastructure is based on the following core basic functionalities, as shown in Figure 2:

- STORAGE: maintaining a large scale entity repository which can ensure the persistent association between a unique entity identifier (ENS-ID) and the corresponding entity.
- MATCHING: mapping any arbitrary description of an entity to its global ENS-ID.
- ACCESS: providing services (i.e. interfaces, APIs) to make ENS identifiers searchable and easily retrievable by humans and machines.
- RESOLUTION: given an ENS-ID in input providing a short description (i.e. entity profile) about the identified entity in output.
- LIFECYCLE MANAGEMENT: supporting few basic operations like entity creation, merging, splitting to ensure the lifecycle management of the ENS identifiers in the system.

By providing a technical infrastructure for the registration and management of global identifiers for use on digital networked environments, the ENS has many features common to existing persistent identifier systems. First of all, the main goal of the ENS is to store the persistent association between a string of characters (the ENS-ID) and an entity. Secondly, ENS identifiers are actionable identifiers but are not locators (URLs). Third, the ENS provides a resolver which allows to enter an ENS-ID and access a small set of metadata providing a short description of the corresponding entity. Fourth, the ENS stores identifiers along with a small set of metadata providing descriptive information about an identified referent. This information is returned by the resolution service. The relationships between the entity, the ENS-ID and the metadata description (entity profile) is shown if Figure 3.

In addition, the ENS has some distinguishing aspects. While many persistent identifier solutions have been developed to identify specific kinds of entities (e.g. DOI and URN for digital objects, ORCID and ISNI for authors and contributors), The ENS-IDs are digital identifiers for entities of any type (digital and non-digital entities) like people, institutions, publications, Web pages, events, locations and so on. Another difference concerns the scope of the identification system. The majority of the current persistent identifier solutions were introduced to solve the problem of changes in location or name of the resources on digital networks (i.e. the broken link issue) by maintaining a persistent binding between the identified resource and an online location where the object or a representation of it can be retrieved. The ENS has been developed as a service for enabling the fulfillment of entity-centric approaches for data integration in digital distributed environments, like the Semantic Web. The issue in this second case is distinctly related to global naming and reference rather than to persistent resolution. Finally, the ENS metadata model has not been developed to address semantic interoperability issues (like for example the DOI data model), that is enabling the automatic reuse of information originated in one context in another context, but has been created to enable disambiguation and entity matching within the ENS identifier repository. The ENS metadata model consists of a minimum set of metadata which should be sufficient to uniquely identify the entity and distinguish it from the other stored entities. The metadata are used for making the identifiers searchable and retrievable (search queries are matched on metadata values) and to provide a short description of the identified referent to a user.

From the above comparison it emerges that the ENS has the potential to fill some of the interoperability gaps of the PID landscape even though an evolution of the system is required. As we have stated in the introduction of this paper, one of the main challenges of the modern research infrastructures is not only to allow persistent access and reuse of digital information, but to create a global interoperability environment where data and information can be seamlessly exchanged across disciplines, institutions and services and integrated knowledge can be extracted through an articulated network of connections linking all the relevant entities in the landscape, like for example data to authors, contributors and journal articles, authors to publications, co-authors and institutions, projects to institutions, authors and funding agencies and so on. The value of these connections can be used to provide added-value services like citability, tracking of research output, quality metrics, provenance and many others. One of the major gaps to exploit the value of this connectivity is the lack of interoperability between current PID systems which hinders the possibility of creating and navigating this valuable network and leads to the creation of information islands in a very similar way to what has been described for the Semantic Web. This is not surprising since tailored local PID solutions have been developed with the aim of addressing needs of specific communities without having interoperability purposes in mind. The ENS has been instead designed as an interoperability solution from the beginning. In the next section we will discuss how the ENS can realize the technological infrastructure for addressing the instance-level information integration problem at the core of e-science infrastructures. Some recent crucial modifications and additional functionalities are also presented as part of the evolution of the system toward a novel infrastructure capable of satisfying the three main requirements discussed in Section 3

5. THE EVOLUTION OF THE ENS TOWARD AN INTEROPERABLE INFRASTRUCTURE FOR PERSISTENT IDENTIFIERS

Up to this point, the ENS has been presented as an infrastructure supporting the identification of several types of entities and implementing a sophisticated matching mechanism to allow the reuse of identifiers across independently produced content. However, three additional features need to be addressed by the ENS in order to become a productive interoperability infrastructure for persistent identifiers in e-science.

First of all, the system should not operate as a centralized solution for global persistent identifiers but as an integrating infrastructure federating current persistent identifier solutions to ensure interoperability. It has become clear in the last few years [5] that a unique global identifier solution is not the right answer to the interoperability problem of identifiers. This is because many solutions have been consolidated in some domains (e.g. publishers or institutional repositories) and local tailored systems are difficult to be overcome since they provide services tuned to the specific needs of specific stakeholders. To work as an integrating PID infrastructure the ENS needs to facilitate interoperability between systems already in use and support the development of added value services which can address both specific community needs and cross-boundary requirements. Technically this can be realized through an effective management of mappings between the ENS identifier assigned to a given entity and any other (persistent) identifier for the same entity (alternative ID management service). In this way, an ENS-ID can be viewed as unifying integration service providing a single entry point to multiple alternative identifiers for the same entity. The ENS infrastructure has the basic core service for registering and managing alternative identifiers. All the alternative IDs available for the entity are stored in the ENS registry as part of the entity profile (see Figure 4). The functioning of the alternative ID management service can be understood by performing a simple query for an entity through the search interface of the ENS[12]. For example by entering the keyword <Tim Berners-Lee>, the ENS (through its default resolver) returns a short description of the scientist through its core set of metadata of the entity profile and a list of the alternative identifiers for the searched entity. Figure 5 shows the screenshots for the example query. In the example the alternative identifiers for the target entity (i.e. Tim Berners-Lee) are URLs belonging respectively to dbpedia and freebase namespaces. The "alternative-id" relationship between them and the binding to the ENS-ID of the entity has been established through the matching functionality when structured information about the target entity has been imported from these knowledge bases into the ENS. The matching algorithms implemented in the ENS use the descriptive metadata in input to establish if an ENS-ID has already be assigned to the entity. If the entity has already registered, the import function updates the profile and imports the IDs used in the original sources as alternative IDs. Otherwise a new profile is created and the imported information is used to fill the core metadata of the profile (through vocabulary mapping) including the alternative ID field. The alternative ID management service could be used to map any kinds of alternative identifier including alternative persistent identifiers, like for example,

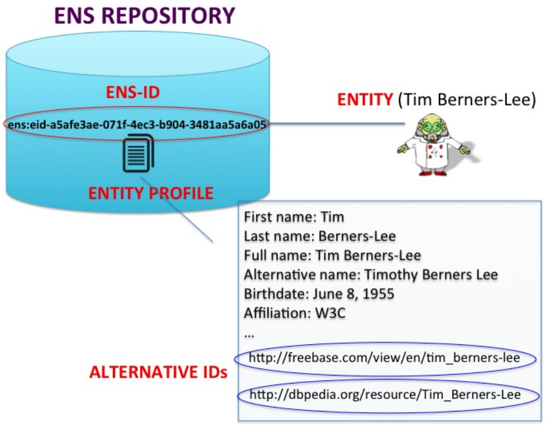
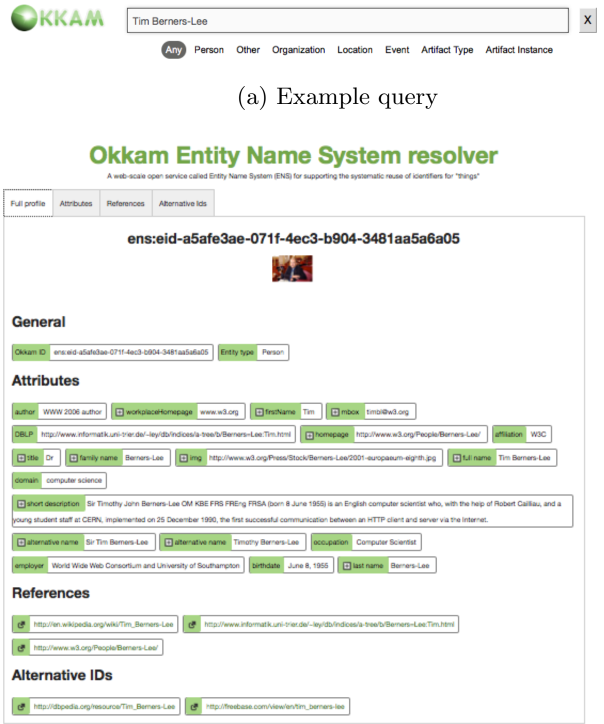

ENS REPOSITORY
ENS-ID
ens:eid-a5afe3ae-071f-4ec3-b904-3481aa5a6a05
ENTITY PROFILE
ENTITY (Tim Berners-Lee)
First name: Tim
Last name: Berners-Lee
Full name: Tim Berners-Lee
Alternative name: Timothy Berners Lee
Birthdate: June 8, 1955
Affiliation: W3C
...
ALTERNATIVE IDs
http://freebase.com/view/en/tim_berners-lee
http://dbpedia.org/resource/Tim_Berners-Lee

Figure 4: ENS Alternative ID Management Service

[12] The search interface is available at `http://api.okkam.org/search/`



Tim Berners-Lee

(a) Example query

Okkam Entity Name System resolver

A web-scale open service called Entity Name System (ENS) for supporting the systematic reuse of identifiers for "things"

ens:eid-a5afe3ae-071f-4ec3-b904-3481aa5a6a05

General

Okkam ID ens:eid-a5afe3ae-071f-4ec3-b904-3481aa5a6a05 | Entity type Person

Attributes

author WWW 2006 author | ⊞ workplaceHomepage www.w3.org | ⊞ firstName Tim | ⊞ mbox timbl@w3.org
DBLP http://www.informatik.uni-trier.de/~ley/db/indices/a-tree/b/Berners=Lee:Tim.html | ⊞ homepage http://www.w3.org/People/Berners-Lee/ | affiliation W3C
⊞ title Dr | ⊞ family name Berners-Lee | ⊞ img http://www.w3.org/Press/Stock/Berners-Lee/2001-europaeum-eighth.jpg | ⊞ full name Tim Berners-Lee
domain computer science
⊞ short description Sir Timothy John Berners-Lee OM KBE FRS FREng FRSA (born 8 June 1955) is an English computer scientist who, with the help of Robert Cailliau, and a young student staff at CERN, implemented on 25 December 1990, the first successful communication between an HTTP client and server via the Internet.
⊞ alternative name Sir Tim Berners-Lee | ⊞ alternative name Timothy Berners-Lee | occupation Computer Scientist
employer World Wide Web Consortium and University of Southampton | birthdate June 8, 1955 | ⊞ last name Berners-Lee

References

http://en.wikipedia.org/wiki/Tim_Berners-Lee
http://www.informatik.uni-trier.de/~ley/db/indices/a-tree/b/Berners=Lee:Tim.html
http://www.w3.org/People/Berners-Lee/

Alternative IDs

http://dbpedia.org/resource/Tim_Berners-Lee
http://freebase.com/view/en/tim_berners-lee

(b) Seach output

Figure 5: ENS search interface screenshots

referring to our previous example, the Scopus ID and the ORCID ID for Tim Berners-Lee (if available). This mapping would enable a first level of interoperability between the two identification systems allowing to identify (and access) two islands of information in the corresponding systems which refer to the same entity and create a bridge between them. Going back to our example, entering a Scopus ID one can find the alternative ORCID ID and by resolving this ID, access to information about the target entity. In the case of digital objects, the alternative identifiers can be used to get alternative access to the resource on different servers as well as related information. For this purpose, a redirect service, based on the alternative IDs associated to the ENS-ID, has been recently developed which allows users to resolve the ENS-ID into third-party data sources[13]. For a given ENS-ID, the service allows to get a list of resolvers and redirect to a selected resolver. It should be noted that the ENS approach for managing alternative identifiers differs from that proposed by the APARSEN Interoperability Framework. In the APARSEN framework the co-reference between alternative identifiers is provided directly by content providers and this mechanism allows to create a linkage between previously disconnected resources (see footnote 9). On the contrary, the ENS alternative ID management service connects the alternative identifiers to the profile of the identified entity and therefore links them to the unique ENS-ID for that entity. In this way, the ENS-ID works as the glue for bridging all the alternative IDs referring to the entity. Any of these IDs (in use in different systems) can be used to interrogate the ENS and retrieve the corresponding unique ENS-ID which in turn gives access to all the alternative IDs of the profile. Through the alternative IDs, alternative ways of access to the resource or information about the resource are enabled, empowering the cross-boundary integration and mash-up of data. Moreover, a profile can be updated with additional alternative identifiers across time as the entities named in different sources are matched and aligned with the ENS identifiers via a process of automatic entity matching.

A second aspect deals with persistence. In [4] we have discussed the evolution of the ENS to a persistent ENS through the separation of the ID (e.g. `peid?8af7c50f?f072?4384?905b?03875c341863`) from the resolver (`http://www.okkam.org`). This introduces a level of indirection between the identifier and its referent and ensures the persistent binding between them. By default, the ENS-ID is combined with the ENS default resolver and its resolution returns a small set of metadata (included in the ENS entity profile) related to the identified entity. The real potential of separating the token id from the resolver rests on the possibility of associating the same ID to multiple resolvers, enabling a mechanism of multiple resolution. Different actors can create or reuse persistent ENS-ID (PEID) for entities of interest using the ENS and through their local resolvers enable precise (and long-term) access to information they store (see Figure 6 extracted from [4]). While ID management is addressed by the ENS, information management, including persistence of the content, and reliable resolution (excluding the default resolution service provided by the ENS) is managed by content providers, in line with the main assumptions of the APARSEN interoperability framework for PIDs but also addressing the requirements of the linked data community. The ENS PEIDs can be reused as part of Cool URIs allowing Linked Data users to create URIs resolvable to any information source they like. At the same time, persistent identifiers users can reuse the same PEIDs to identify information objects and resources managed by trusted institutions which ensure their persistent access and association to a physical location. Due to this change of paradigm, the ENS differs from a centralized authoritative service for minting and resolving global identifiers, allowing to every one the reuse of the ENS-IDs to create persistent identifiers (through domain resolvers) or Cool persistent URIs (through the web service resolution mechanism). The last point is important since several initiatives[14] have highlighted the need to develop a co-ordinated solution to identifier issues across the PID and the Linked Data community (as stated for example in the Den Haag Manifesto[15]). The recent improvement of the ENS may offer such a solution, enabling data creators and curators to combine the technical strengths and opportunities of the (Semantic) Web vision with the organizational, economical and social requirements legitimately raised by the PID community and stakeholders. This has a strong impact on the development of services to support the integration of information across sources since it opens the door to new forms of interactions between open structured data published on the Web and content stored by more

[13] More information is available at `http://community.okkam.org/`

[14] For example, the Persistent Object Identifiers seminar at The Hague in June 2011 and the Links That Last workshop in Cambridge in July 2012

[15] available at `http://www.knowledge-exchange.info/Default.aspx?ID=462`

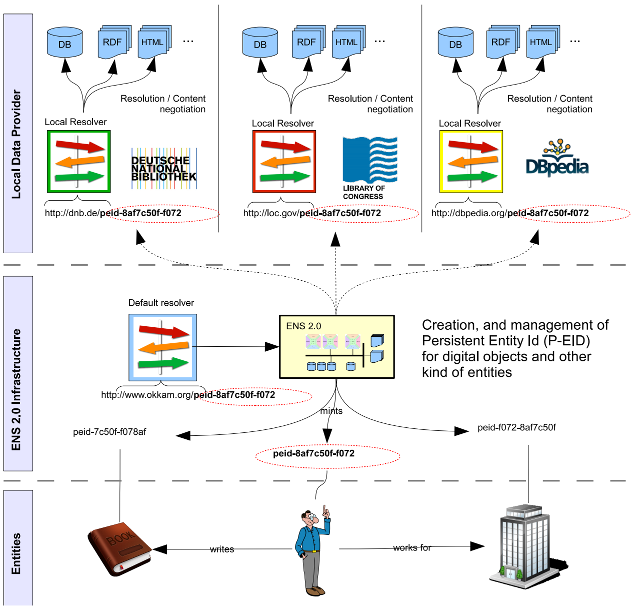

Local Data Provider
DB
RDF
HTML
...
Resolution / Content negotiation
Local Resolver
DEUTSCHE NATIONAL BIBLIOTHEK
http://dnb.de/peid-8af7c50f-f072
LIBRARY OF CONGRESS
http://loc.gov/peid-8af7c50f-f072
DBpedia
http://dbpedia.org/peid-8af7c50f-f072
ENS 2.0 Infrastructure
Default resolver
ENS 2.0
Creation, and management of Persistent Entity Id (P-EID) for digital objects and other kind of entities
http://www.okkam.org/peid-8af7c50f-f072
mints
peid-7c50f-f078af
peid-8af7c50f-f072
peid-f072-8af7c50f
Entities
BOOK
writes
works for

Figure 6: Multiple Resolution in the ENS

traditional cultural heritage institutions.

The third aspect deals with vocabulary mapping. Different persistent IDs may be associated with different vocabularies used to represent the identified resources. If a mapping among them is available, information structured according to a given schema and retrievable thanks to a given ID can be directly re-used to integrate or update the information of another source adopting a different schema to represent the same entity. Therefore, in order to support semantic interoperability across services and communities, the ENS should provide an extensive mapping of vocabularies and schemes adopted in different PID domains. A service, called OKKAM Synapsis[16], is currently under development to automatically compute the mappings between terms in controlled vocabularies and ontologies toward the ENS core set of metadata. Synapsis is designed as a Web application to support a community-driven effort in the collection and maintenance of mappings. Through the application, a user (human user or API user) can search mappings for a given property by using different filters (e.g. author, status, date), find clusters of mappings for all the registered properties, propose new mappings (which then can be accepted by the administrator of the service) and edit or rate existing mappings (i.e. add comments and manually evaluate mappings by classifying each mapping into one of different categories). While in the APARSEN Interoperability Framework semantic interoperability is addressed by proposing a common ontology which should be used by content providers to expose their data in a common way, the ENS approach focuses on the alignment of different vocabularies through ontology mapping. This has the advantage that users can maintain their own vocabularies and ontologies, without the need to restructure their content according to a new model. The mapping of vocabularies allows supporting the building of crosswalks between them and can be extended to include an indefinite number of vocabularies.

6. BUILDING ADDED VALUE SERVICES ON TOP OF THE ENS INFRASTRUCTURE

A number of added value services can be built on top of the interoperability layer provided by the ENS infrastructure and usable by other systems or infrastructures. We describe some examples.

1. **GLOBAL RESOLUTION SERVICE:** Based on the ENS redirect service described above, a global resolution service can be implemented, which determines the appropriate resolver for a given PID. Moreover, if alternative IDs are associated with the searched PID, the service returns alternative resolvers to access the identified resource via alternative routes.

2. **METADATA ENTITY IDENTIFICATION SERVICE**: This service allows assigning unique identifiers to entities named within the metadata of other resources. For example, if the metadata of a journal publication include author information, the system allows assigning a unique ID to the author which can be an istantaneously generated ENS ID if the entity has not been registered in the repository before, or can be selected among the IDs available in the entity profile if the entity matches one already stored in the system.

3. **METADATA EXCHANGE SERVICE**: By linking a PID to alternative IDs, the ENS interoperability layer can be exploited to develop services for automatic exchange of metadata across systems using different identification solutions. For example, given a PID for an author (e.g. an ORCID ID), the service provides the link to external sources of information (e.g. Scopus, ResearcherID, arXiv) where information about the same author can be found and automatically imported into the original author profile. This can be done thanks to the mapping between the corresponding vocabularies provided by the ENS interoperability layer (via the Synapsis service).

4. **IDENTITY LINKAGE SERVICE**: When a PID for an entity (e.g. an author) is entered, the service returns all the entities related to that entity belonging to a certain entity type (like for example all the author's publications) and allows to navigate the entire chain of links connecting the identified entity to all the related entities (e.g. starting from the PID of a dataset it is possible to go back to the contributors, the related publications, the research projects and so on). Semantic Web technologies provide a possible solution to implement this service. Metadata from different sources can be represented as RDF assertions about resources identified by unique IDs. The ENS interoperability layer offers two unifying elements to integrate data from different sources of metadata: the unique global ENS IDs and their "same-as" relationships with alternative IDs and the vocabulary mappings.

7. CONCLUSIONS

[16] http://api.okkam.org/synapsis/

Interoperability between persistent identifiers is a critical concept for enabling the development of fully-integrated services for research e-infrastructures in order to improve circulation, transfer and access to integrated scientific information and promote cross-boundary collaboration and competition. In this paper we propose a scalable infrastructure to allow current persistent identifier solutions to interoperate and provide integrated access to multiple heterogeneous sources. The proposed infrastructure is based on the OKKAM Entity Name System and implements three main technical core functionalities 1) the management of coreferences among PIDs (alternative id management service) ; 2) the assignment and management of global Persistent Cool identifiers; 3) the mapping of vocabularies across PID domains. Beyond the technical requirements, the implementation of the system will add value to the PID systems only if a governance layer is agreed among them. Therefore, effort is currently dedicated to create the social and organizational support among the relevant stakeholders to transform the ENS into a public open infrastructure for PID interoperability maintained (but not owned) by a Trustee monitored by a board of protectors according to a Trust agreement. As a first step to increase the trust and community support around the ENS infrastructure, we are currently working to propose the ENS interoperability services as part of the offerings of the APARSEN Virtual Centre of Excellence [8] that brings together a diverse set of stakeholders, researchers and practitioners in digital data and digital preservation.

8. REFERENCES

[1] APARSEN D22.1: Persistent identifiers interoperability framework. `http://www.alliancepermanentaccess.org/wp-content/uploads/downloads/2012/04/APARSEN-REP-D22__1-01-1__9.pdf`, 2012.

[2] APARSEN D22.3: Demonstrator set up and definition of added value services. `http://www.alliancepermanentaccess.org/wp-content/uploads/downloads/2014/02/APARSEN-REP-D22_3-01-1_0.pdf`, 2013.

[3] APARSEN D4.1: Conceptual model of interoperability. `http://files.figshare.com/1239137/D4.1_Conceptual_Model_of_Interoperability.pdf`, 2013.

[4] B. Bazzanella, S. Bortoli, and P. Bouquet. Can persistent identifiers be cool? *IJDC*, 8(1):14–28, 2013.

[5] P. Bouquet, B. Bazzanella, M. Dow, and R. Riestra. DIGOIDUNA FINAL REPORT: Digital Object Identifiers and Unique Author Identifiers to enable services for data quality assessment, provenance and access. `http://cordis.europa.eu/fp7/ict/e-infrastructure/docs/digoiduna.pdf`, 2011.

[6] P. Bouquet, H. Stoermer, and B. Bazzanella. An entity name system (ens) for the semantic web. In *ESWC*, pages 258–272, 2008.

[7] P. Bouquet, H. Stoermer, C. Niederée, and A. Mana. Entity name system: The back-bone of an open and scalable web of data. In *ICSC*, pages 554–561, 2008.

[8] D. Giaretta and all APARSEN partners. APARSEN D11.4: Virtual centre of excellence development. `http://www.alliancepermanentaccess.org/wp-content/uploads/downloads/2014/02/APARSEN-REP-D11_4-01-1_0.pdf`, 2014.

[9] A. Hayes, S. Mann, A. Aryani, S. Sabine, L. Blackall, P. Waugh, and S. Ridgway. Identity awareness and re-use of research data in veillance and social computing. In *Proceedings of The IEEE International Symposium on Technology and Society (ISTAS)*, 2013.

[10] T. Weigel, M. Lautenschlager, F. Toussaint, and S. Kindermann. A framework for extended persistent identification of scientific assets. *Data Science Journal*, 12, March 2013.

[11] J. Wood, T. Andersson, A. Bachem, C. Best, F. Genova, D. R. Lopez, W. Los, M. Marinucci, L. Romary, H. V. de Sompel, J. Vigen, P. Wittenburg, D. Giaretta, and R. L. Hudson. Riding the wave - how europe can gain from the rising tide of scientific data. final report of the high level expert group on scientific data. a submission to the european commission. `http://cordis.europa.eu/fp7/ict/e-infrastructure/docs/hlg-sdi-report.pdf`, October 2010.

A Novel Metadata Standard for Multimedia Preservation

Walter Allasia
EURIX
26, via Carcano
10153 Torino, IT
+390112303729
allasia@eurix.it

Werner Bailer
JOANNEUM RESEARCH
Steyrergasse 17
8010 Graz, AT
+433168761218
werner.bailer@joanneum.at

Sergiu Gordea
AIT
Donau-City-Strasse 1
1220 Wien, AT
+43505504274
sergiu.gordea@ait.ac.at

Wo Chang
NIST
100 Bureau Drive
Gaithersburg, MD, US
+13019753439
wchang@nist.gov

ABSTRACT

This paper introduces the motivation and describes the model of a novel metadata standard for the exchange of preservation metadata of multimedia content. The model is being standardised as the MPEG Multimedia Preservation Application Format (MP-AF), addressing the specific issues related to the preservation description information of audiovisual contents. Several standards for expressing metadata in digital preservation are available and have been taken into consideration in the paper and in the work done in MPEG. However, none of them is able to cover all the needed aspects related to the preservation of audiovisual content. Audiovisual files are in most cases containers and are usually made up of several tracks carrying audio data, video data and specific time-based metadata. In order to be able to perform the opportune preservation actions (among others planning and format migration), several kinds of information must be kept alongside the audiovisual contents. Within this context a standardised representation for these structures and metadata is needed. Information such as quality description or fixity at frame level is required for ensuring long term access to visual content. Without using a standardised interface it is hard to guarantee a faithful rendering of encoded information while exchanging contents between different repositories, either internally or with external institutions. This paper describes the work done so far within MPEG for defining a standard metadata model which covers the identified missing parts and gaps regarding the acquisition of digital preservation description information.

General Terms

Preservation strategies and workflows, specialist content type, case studies and best practice.

Keywords

Digital Preservation, Multimedia Metadata, Preservation Description Information, OAIS, Audiovisual Content Preservation, Standard.

1. INTRODUCTION

In the last decade, many research projects at national as well as international level have investigated solutions for preserving audiovisual content. Projects such as the Presto family starting in 1999 have deeply studied the preservation and storage of professional audiovisual contents. Special attention has been given to broadcasting environments that suffer from the obsolescence of audiovisual contents, wrappers and carriers [1,2,3,4]. Even if many formats were available for describing general purpose information, it was clear that many others were left out of the standards and kept in custom structures [3]. As an example, the quality information acquired during the digitisation process (conversion from the analogue audio visual carrier towards a digital representation and support) have been stored in specific structures by broadcasters and audiovisual archives [3].

Many organisations collecting various types of multimedia content, such as archives, libraries, museums, etc. already have digital preservation systems in place. These organisations often have the need to exchange multimedia assets and related metadata, for example:

- to exchange assets between preservation systems/repositories within the organization or with related organizations,
- to change/upgrade their preservation systems,
- to exchange content with service providers, or to
- provide preservation services for other organisations.

When they exchange multimedia assets, they need to include preservation metadata that enables the receiving organisation both to assess the integrity and fidelity of the assets it receives and to establish a baseline for its own curation and use of the assets. In addition to the metadata described above, the receiving organisation also needs information about any preservation processes the assets have undergone, including descriptions of the outcome of such preservation processes. The description may include metadata about content, structure, and quality, as well as technical, historical and editorial information, and information about property and use rights and conditions. A standard is needed that defines the content and format of multimedia preservation description information (MPDI), in order to facilitate interoperability between preservation systems, ensure accurate understanding of the resources' exchanges, and reduce the risks of corruption both in the exchange and thereafter.

These issues have been strongly pointed out in the UNESCO Vancouver declaration, written during the UNESCO conference "Memory of the World in the Digital age: Digitization and Preservation" held in Vancouver in 2012 [5]. The overall document stresses the need and the importance of digital preservation. The following quotation taken from the last part of the document reinforces the need of a metadata standard for multimedia preservation:

”Recommendations to industry: ...

a. *ensure long-term accessibility to digital information;*
b. *adhere to descriptive standards and recognized metadata standards to enable the creation of trusted digital repositories. ...”* [5]

We can easily recognise that the work presented in this paper responds to what (b) is motivating and UNESCO is stressing the importance of metadata standards for preservation in order to enable the creation of preservation archives, or more precisely, the trusted digital repositories.

This paper describes the work on a metadata model for multimedia preservation metadata within the standards ecosystem of MPEG [6]. MPEG is an ISO/IEC working group defining standards for coding moving pictures and audio. Over the years the work of MPEG has broadened to include metadata over the lifecycle of multimedia items. Thus a standard for preservation description information of multimedia items complements these efforts. Within MPEG, the work on multimedia preservation is done in the context of application formats, which are standards composed of subsets of different MPEG technologies targeting a specific application scope, and extending them with existing technology from outside MPEG or new technology if needed. The preservation metadata standard is thus named Multimedia Preservation Application Format (MP-AF).

The metadata model proposed by MP-AF is presented in this paper, which is organised as follows: Section 2 introduces the motivation and foundation of the work done while Section 3 discusses related work. Section 4 presents the Multimedia Preservation Description Information (MPDI), in accordance to the OAIS model and definitions [7]. The metadata model describing the multimedia preservation metadata is presented in Section 5, and its relation to other data models is discussed. Section 6 wraps up the conclusion of this work and points out upcoming activities.

2. MOTIVATION

There is a range of different organisations that are in charge of preserving multimedia content, from dedicated audiovisual archives with a mission for preserving the collection over archives of broadcasters or media production companies, for which the archive must primarily serve their production workflows, to libraries or museums that may only have some multimedia items among their collection. All of them have the need to perform processes such as digitisation or migration of the multimedia content. These processes are increasingly automated within large organisations or outsourced to service providers by organisations that cannot afford to have the infrastructure and knowledge in-house. Performing and documenting these preservation workflows requires the ability to represent detailed preservation description information in an interoperable way.

The use cases addressed by MP-AF include (partial) preservation workflows where metadata created or needed in the preservation process is exchanged between different systems or organisations. One example is a preservation workflow including regular fixity, integrity and quality checks of broadcast content, performed by different systems. Another example is a migration process outsourced to a service provider, which includes determining the need for migration to another format, choosing the parameters of the target format and performing quality checks to ensure that the result of the migration is free or errors and is a complete representation of the source. Another use case involves content being exclusively licensed to another company (under certain restrictions, e.g. territorial) and making sure to identify the versions affected by the contract, including different technical formats for different distribution channels.

A different group of use cases addresses cases where content is deposited with an archive for preservation. For example, a national library may have the mission to preserve all music recordings, and receives the masters from the record companies, including the metadata for each song and the collection, as well as further artwork related to the production. The carrier may become obsolete and digitised/migrated to another carrier or file storage, keeping the grouping of the different objects in the submission and their relations.

PREMIS is nowadays the (de facto) standard which is used by many national libraries and archives for aggregating and preserving metadata required for ensuring long term access to digital content. Key concerns are related to the renderability, understandability and identity of digital objects with the passing of time. Repositories that store the digital items related metadata, must ensure their consistency over time. The standard makes no assumptions about the preservation strategies, technologies and storage systems. It is meant to be used on any type of digital content in any available encoding (i.e. file format). PREMIS defines the dictionary of preservation metadata elements, but not the structure of the description resp. the metadata container. It thus needs to be embedded in some container structure, for example, METS or MPEG-21 DID. This way, one can aggregate more complex archiving structures related to book collections, movie series, photo exhibitions, etc. (cf. [8]).

When using the PREMIS standard in a concrete application scenario, it is soon observed that different enhancements are required to address particular needs of a given preservation context [8]. In particular, the following issues have been recognised in the context of preservation metadata for audiovisual content.

Compatibility with standards in use. MPEG standards are widely used by broadcasters and audiovisual archives. The information relevant for preservation purposes is partly covered by descriptive and technical metadata standards already in use. Compatibility with these formats eliminates the overhead required for mapping and transforming existing metadata to PREMIS representation and may ease acquisition of preservation related metadata during content creation (e.g., collection of timing and location metadata with digital cameras, metadata acquisition at digitisation time). These compatibility issues do not only concern metadata formats, but also container formats like the MPEG Professional Archive Application Format (PA-AF) [9].

Enhanced support for modelling hierarchical, complex structures and descriptions. A collection is a common unit of work in digital libraries and archives. Collections may be aggregated in hierarchical structures by using different criteria. Multimedia content is often the result of a long and complicated creation process, reusing material from a multitude of sources, each with their specific properties, provenance and rights. For example, it is popular nowadays to have long TV series organised in seasons and episodes, including versions translated in different languages. Motion pictures may be released in a number of localised and age versions, with different audio formats, in different 3D technologies etc. Moreover, the file formats for encoding this content is a container itself carrying bitstreams of different types of data: audio, video, subtitles, etc. Over its lifetime, the content may need to be migrated due to obsolescence of the original formats. For ensuring the long term access to the content by respecting copyrights and ownership, it is mandatory to preserve descriptive and technical metadata at each level of aggregation.

Support for time-based metadata. The existence of a temporal dimension is an inherent property of audiovisual content. For many types of metadata, it is crucial to have them on a detailed temporal granularity, for example, per shot. This includes descriptive and technical information, which may differ as the shots may be recorded with different technologies. In types of productions that rely heavily on the reuse of material (e.g., news), each shot may come from a different source, having its specific provenance and rights metadata. Due to the potentially long duration of a content item and its large file size, it is also important to have quality and fixity metadata on a fine temporal granularity in order to locate and potentially repair problems in later steps of a preservation workflow.

Defining the metadata container. The PREMIS standard does per se not specify the metadata container, for example, for the creation of submission, archival and dissemination packages as defined in the OAIS standard. As the choice of the container is left to the implementation, there are no built-in mechanism for ensuring the referential and data integrity of the package. Consequently in the case of broken packages there is no mechanism defined for verifying which parts of the package are not corrupted and can still be used properly in preservation processes.

MP-AF aims to address these issues by defining a specification that provides solutions for these gaps. Compatibility with PREMIS has been taken into account in the design of the standard, and mapping is intended to be straight forward for overlapping parts of the specifications. Moreover, the MP-AF representation takes into account additional issues related to the encoding the metadata in different languages using alternative scripting variants and extendable semantics of the core elements by using controlled vocabularies. By standardising the format of the metadata container and referencing within of the information package a better support for implementation of preservation workflows and outsourcing of preservation services can be provided.

3. RELATED WORK

While an abundance of metadata standards and formats for describing multimedia content exists, this is not the case for the description of material properties, tools and processes for preservation of audiovisual content. Preservation metadata is a relatively new concept, and preservation metadata models emerged quite recently in the digital library domain. The most important of these models is PREMIS, a model proposed by the US Library of Congress [10]. It defines a high-level data model and a set of properties for each of the entities in the model. There are five semantic units (classes) in PREMIS as shown in Figure 1. An XML representation exists and an OWL (Web Ontology Language) representation has recently been proposed [11].

One issue related to the modelling of multimedia content is the assignment of rights to an Agent, which is different from general the licenses or contracts related to the object or intellectual entity. There is a second issue, as the digital object and the intellectual entity are considered at the same level. Furthermore, rights represented as an association class between object and agents are expressing the “access control” instead a full “contract” that is usually applied in multimedia environments. Subclasses of Object are File, Bitstream and Representation that are suitable for multimedia content as well. However, the typical hierarchy of representing levels of multimedia content between the work and a specific bitstream (as e.g. commonly used in the broadcasting domain, see the description of EBU CCDM below) are not directly supported. This concerns in particular the issue that a multitude of versions of multimedia contents exist, and regular migration between different technical formats (potentially as a lossy process) is a common issue.

The National Library of New Zealand (NLNZ) has defined a metadata model largely based on PREMIS [12] by adding extensions and addressing some implementation issues. The AIP is made up of the digital object, the technical metadata required for technical preservation of the object (using Rosetta DNX) and the descriptive metadata required for discovery and asset management.

Another approach to represent the provenance of digital objects, related events and agents was recently proposed within the provenance model developed by the W3C [13]. This has evolved in the context of open data initiatives, in order to track the activities that created and modified data published on the web. The core of this data model is represented in Figure 2, where three elements were identified, namely the Agent, Entity and Activity. In addition to the data model, the PROV family of specification defines different serializations of the model, including XML and RDF/OWL.

A model for the authentication of digital resources and representing the steps in the process that impact authenticity has been proposed on [14] and later refined in [15]. When dealing with multimedia content, the implementation of PREMIS elements in MPEG-21 Digital Item Declaration (DID) containers has been proposed in [16]. However, the link between the PREMIS descriptors and the MPEG-21 structures is rather loose, not fully leveraging the potential of both technologies. A similar approach is used by D2D, which is another MPEG-21 based representation for preservation metadata [17]. The core of the model is based on MPEG-21 DIDL and a set of specific descriptors was defined, which hold the various types of preservation metadata. Many of the descriptors are specified in a very generic way, using a key/value representation.

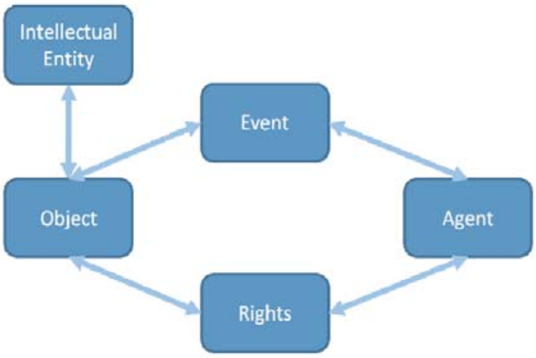


Figure 1: Data Model of PREMIS (from [10]).

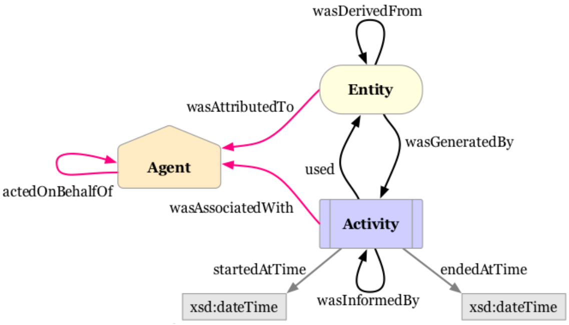


Figure 2: The conceptual overview of the PROV data model and related properties (as presented in [18]).

Within the PrestoPRIME project, a specific data model has been defined in order to set up an OAIS archive of audiovisual contents [19,3]. PrestoPRIME faced the problem of managing AV contents on a long term basis, therefore the OAIS specifications had to be adapted to deal with specificities of multimedia contents. In the PrestoPRIME model, the EditorialEntity has one or more Representations which have associated Files made up of Bitstreams. This data model is quite powerful and covers several requirements of MP-AF. The model uses METS as wrapper and includes elements from DublinCore, PREMIS, MPEG-7 and MPEG-21, as well as some custom extensions, such as DNX.

EBU CCDM [20] is a conceptual data model for audiovisual content and related entities from the broadcasting domain. The main entities are Intellectual Property Rights, the Production Order, the Sales Order, the Editorial Objects, the Asset (EditorialObject with associated IPR), the Manufacturing Object, the Publication Event, the Media Resource Object. Although not specifically designed for preservation, several entities overlap with preservation metadata models.

The same holds for the MPEG-7 Detailed Audiovisual Profile (AVDP [21]), which is part a MPEG-7 metadata standard defined for applications in production and archiving of audiovisual content. This standard covers technical and descriptive metadata, including quality analysis results, but lacks other aspects needed for a preservation metadata model.

Another related technology from the multimedia area is the MPEG Professional Archive Application Format (PA-AF) [9]. Like MP-AF, it is an application format combining different MPEG technologies for use in the archival domain. However, its focus is on specifying a virtual structure for packaging multiple items into a single file in order to preserve them together in a platform independent way. The resulting file conforms to the MPEG-21 file format, while providing only very basic metadata support. It is thus complementary technology to MP-AF and the two technologies can be used together, as described in Section 5.4.

4. THE MULTIMEDIA PRESERVATION DESCRIPTION INFORMATION (MPDI)

The data structures laying behind the definition of the preservation objects are presented in Section 5.1. In the following we present the most important concepts formalizing the representation of the information used for multimedia preservation purposes. These concepts were identified within the scope of the MPDI requirements document, but a revised definition is used within the proposed standard. The model is partly inspired by the PREMIS and partly by the related work in preservation projects (see also Section 3). However, it takes advantage from the complete representation of the preservation information package and includes semantic elements attached to each level in the hierarchical structures.

The following MPDI concepts define elements relevant for multimedia preservation and used in preservation processes.

Provenance documents the chronology of events regarding the creation, modification, ownership and custody of resources, such as who produced it and who has had custody since its origination. It provides information on the history of the multimedia content (including processing history).

Context describes the circumstances that resulted in the production of the resource and how the preserved resource relates to other relevant resources. For example, it may describe why and how the resource was created, it may indicate from which resources the current one was derived, or it may specify the relationship to other resources available in the package.

Reference represents the information that is used for identifying and addressing the multimedia content and related resources. It uses one or more identifiers, or systems of identifiers, by which the resources may be uniquely and persistently identified. Reference information supports the linkage of identical or related resources that might be stored in separate repositories. These repositories may use different mechanisms for identifying resources (e.g. using different standards for representing local identifiers).

Quality encompasses information related to the qualitative or quantitative measurements describing a given resource. It supports reasoning and evaluation of how good the resources have been preserved. The quality assessment should document any modification or transformation applied to the content, as well as the processes that produced them.

Fixity encompasses the information ensuring that resources (as described by their properties) are not altered in an undocumented manner. This information is also used to verify the integrity of digital items.

Integrity represents the state of an entity (e.g. digital item) indicating the quality of being complete. It can be proven by verifying the presence of all required parts/components in an unaltered (i.e. not modified) state.

Authenticity encompasses information that enables an agent to verify if an object is correctly identified and free from (intentional or accidental) corruption (i.e. it is capable of delivering its original message). The agents that issue statements about authenticity must also be correctly identified. While integrity is only on a technical level, authenticity is concerned with the object not being tampered with. Assessment of authenticity may require information related to different representations of the same work, while integrity refers to a specific representation. For example, the digitisation of an analogue copy cannot be automatically checked for integrity but is still preserves the authenticity of the content. Similarly, transformations from SIP to AIP or from AIP to DIP preserve authenticity, but not necessarily integrity.

Rights encompass information concerning legal, regulatory or contractual provisions that affect ownership, control, access or use of resources insofar as they impact the long term preservation (e.g. intellectual property, copyrights, privacy, etc.). Actions or events in the preservation of resources need to respect such rights.

5. MP-AF DATA MODEL

The MP-AF data model represents metadata for the preservation of a variety of media, such as images, graphics, video, animation, sound and text, and combinations of these. The definition of these elements/classes follows the goal of maximizing interoperability and maintaining compatibility with existing preservation data models. This should facilitate the adoption of MP-AF model among organizations that already use compatible models, at least for data exchange purposes, such as the migration between preservation systems (for software or hardware upgrade for example) or for exchange between repositories.

The MP-AF data model is defined for representing the Multimedia Preservation Description Information (MPDI) needed for discovering, accessing and delivering multimedia resources.

The specification of MP-AF contains three main components. The first is a high-level data model, specifying the top-level entities and their relations. The second part concerns the specific metadata structures for the different types of preservation metadata covered by MP-AF, modelled as descriptors. Whenever possible, these definitions make use of existing metadata standards, i.e., the specification reuses parts of MPEG-7, MPEG-21 and also defines extensions to existing metadata standards (e.g., MPEG-7). The third part (not described in detail in this paper) defines a core set of technical and descriptive metadata that is required to ensure minimum interoperability between preservation systems. A serialisation of the MP-AF data model using XML Schema has been specified.

5.1 Data Model Overview

The central entities in the model are those representing multimedia content. They are designed to be compatible with the MPEG-21 Digital Items, which hold metadata and references to the actual essence. In order to align the proposed model with other ones uses in the media industry four levels of specialisations are defined.

A Preservation Object combines information describing the intellectual and artistic attributes of a Work together with Digital Items that encode the Work. It includes technical, descriptive and preservation metadata and any other information needed to ensure consistent and reliable access to the Digital Item(s) over time. An Asset is a specialisation of Preservation Object aggregating a description of the owner and the owner's rights. These rights are exploitation rights that are different from the usage rights of a Digital Item. This is aligned with the definition of an Asset by the Society of Motion Picture and Television Engineers (SMPTE), which defines assets as being content with associated rights. Preservation Objects may be recursively nested in order to express groups of objects, which constitute a Preservation Object themselves (e.g., tracks of an audio CD vs. the entire CD). In contrast, Groups are explicitly containers of Preservation Objects and not an Preservation Objects themselves (i.e. it a logical grouping such as a broadcasting series).

A Representation is a specific and complete manifestation of the Work. Representations may differ in terms of technical or descriptive properties while sharing the same intellectual and/or descriptive attributes of the Work (e.g. different performances of the same Work, low vs. high definition representations of a movie). A Representation aggregates the whole set of Essences plus any additional metadata needed for a complete presentation of a Work.

Essence is a manifestation of a Work or part of a Work. It refers to the metadata needed for correctly rendering media content including all associated Components.

The Component is the entity holding specific technical metadata supporting the handling of the media resource referenced by a Media Locator (reference or identifier of a storage media volume, Item or part of an Item). Components can be Files or Bitstreams.

Operators are persons, organisations or systems that can be instantiated in form of Agents (persons, organizations) or Tools (hardware devices, software applications). They are involved in a certain Activity with a specific role. Different Agents may have relations to each other. An Activity is a preservation action performed on at least one Digital Item or Component. The activity is carried out by one or more Operators known to the preservation system.

The complete data model is shown in Figure 3. The relations in this diagram are of the following types: inheritance (the entity is a specialization of a more general type inheriting the parent's attributes), composition/aggregation (the entity aggregates other entities) or associations.

The data model contains entities marked with the <<Metadata>> stereotype, which correspond to the metadata types specified in the MP-AF requirements. These entities might correspond to a single or a set of the descriptors in a concrete representation of the model. The use of the metadata types on specific entities of the MP-AF data model is listed in Table 1.

Table 1: Overview indicating the relations of preservation information concepts to content entities.

	Preservation Object	Representation	Essence	File/ Bitstream
Provenance	Yes	Yes	Yes	
Context	Yes			
Reference	Yes	Yes	Yes	
Quality	Yes	Yes	Yes	Yes
Integrity	Yes	Yes	Yes	
Authenticity	Yes	Yes	Yes	
Fixity	Yes	Yes	Yes	Yes
Rights	Yes	Yes	Yes	

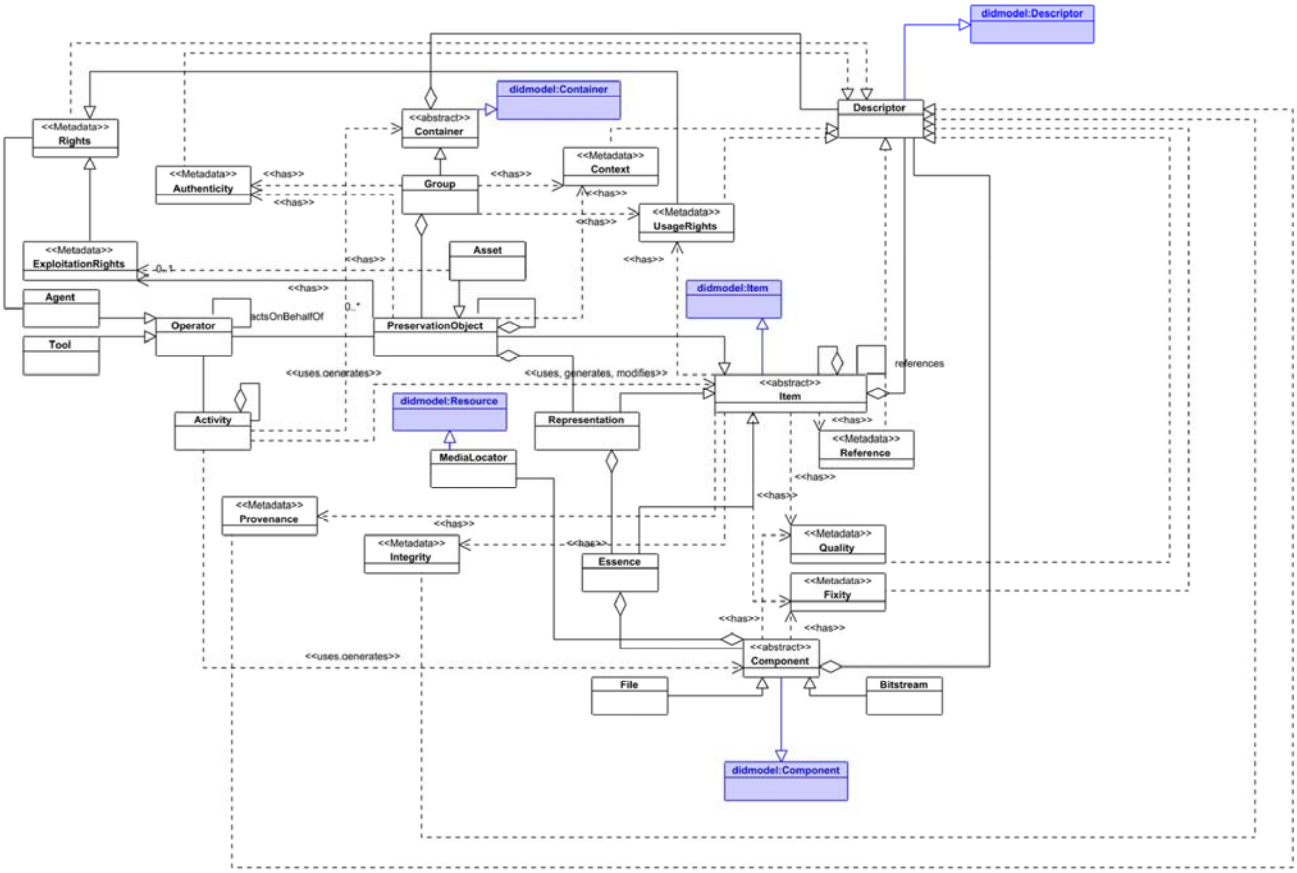


Figure 3: Core MP-AF data model. Entities from MPEG-21 DID are highlighted.

5.2 Structures for Specific Metadata Types

This section describes the concrete representations used for the specific types of preservation metadata. Provenance metadata, item identification and description are supported by MPEG-21 DII and DID. The structure based on MPEG-21 provides sufficient capabilities for item identification, supporting multiple identifiers, qualified by type.

Basic support for structural relationships between the items is provided by the Digital Item structure. Relationships are used to specify alternative identifiers of digital items and their nature/type (e.g., ISBN vs. barcode vs. bookshelf ID). MP-AF adopts MPEG-21 Digital Item Semantic Relationships [22] for expressing relations between Digital Items.

The preservation processes applied to Digital Items and Groups, the Activity and Operator entities are defined as part of the core MP-AF data model. These entities have been defined to ensure maximum compatibility with the corresponding entities in PREMIS, W3C PROV and BPMN (see also Section 4 for further details).

As part of the preservation metadata of multimedia asset, the history of creation and processing steps applied, as well as their parameters are described. This representation thus differs from process model representations including branching and options. The processing log describes what actually happened with a Digital Item, i.e., it is a linear sequence of activities, with the option to add a hierarchy for grouping activities. The descriptions of activities in the model use a set of specific types (e.g., digitisation), with possible further specialisations (e.g., film scanning), in order to improve interoperability between preservation systems. In a similar way, types of tools/devices being operators in these activities are identified. In addition, parameters of tools/devices are represented in a key/value structure, with a set of defined key for the most important properties are specified.

The definition of reference vocabularies is out of scope of MP-AF. However, preferred vocabularies of terms are recommended in an annex of the MP-AF specification where applicable (e.g., the set of quality items being defined by the EBU Quality Control group [12]).

Context can represent information about the purpose of preservation (e.g. project and preservation program). Relationships represent relations between different Digital Items, while Context includes relations to any type of related resource.

For fixity metadata temporally fine-grained checksums are supported. This enables better localization (and thus more efficient repair) of errors in bitstreams.

Integrity metadata comprises of information to index a set of content items, a set of identifiers to be checked and a list of dependencies on other preservation information packages (e.g. collection, and packages of individual episodes). Component-level fixity information as well as fixity of metadata documents/fragments may be included. Format validation results can be represented using the quality descriptors, performing wrapper or bitstream layer checks.

MP-AF can include metadata to support checking of authenticity but cannot ensure the authenticity of the preservation object. MP-AF supports the following information for checking authenticity: information provided by the submitter of the object to the archive (descriptive, rights and provenance metadata), a complete log of the activities related to the preservation object, including technical, organisational and legal activities (based on the process description model) and metadata for comparing representations of the Preservation Object (e.g. fingerprints). As this information can be represented in different descriptors of MP-AF, no specific authenticity descriptor is provided.

MP-AF provides means to represent metadata related to (semi-) automatic quality control of multimedia data. The quality metadata description framework specified in ISO/IEC 15938-5:2003/Amd 5 is used for this purpose.

The description is compatible with the data model defined by the EBU QC group. It is recommended that the metadata refers to the taxonomy of quality control items defined in [12].

For rights metadata, both MPEG-21 Rights Expression Language (REL, part 5) and MPEG-21 Contract Expression Language (CEL)/Media Contract Ontology (MCO) will be supported (MPEG-21 CEL, part 20, and MCO, part 21, are semantically equivalent, but with XML and RDF/OWL representation respectively). REL can be used if it is sufficient to represent the rights situation, i.e., if only usage rights need to be represented. In more complex cases, if media-related contracts need to be documented or when a documentation of the history of the rights situation is needed, CEL or MCO can be used.

5.3 Descriptive and Technical Metadata

The common core set of metadata represents basic information needed to support digital preservation. It includes descriptive metadata by which a Digital Item can be unambiguously identified. It also includes technical metadata (e.g. describing the format) associated with one or more specific Representations, Essences or Components of the corresponding Preservation Object. The common core metadata set should be seen as a baseline profile, which enables a minimal set of knowledge acquisition by the receiver of an MP-AF instance. Additional metadata can be inserted into the MP-AF structure at any place that allows MPEG-21 compliant descriptors.

The common core metadata set is modelled as two specific descriptor types. Descriptive metadata contains a basic set of descriptive metadata elements, represented either as MPEG-7 Creation Information, EBU Core or Dublin Core descriptor. Technical metadata defines basic technical information (for different media types), represented as MPEG-7 Media Information or EBU Core Format descriptors.

5.4 Relation to Other Data Models

The interoperability with other existing data models related to digital preservation has been adopted as a core design principle. The purpose of MP-AF is not to provide yet another metadata standard, but the most interoperable and complete metadata standard for describing the preservation information needed in professional audiovisual domains. Three data models have been selected as the most adopted in the current practice of audiovisual archives, and therefore as mapping targets: PREMIS [10], W3C PROV [13] and EBU CCDM [20]. These mappings for MP-AF to already implemented data models are useful for understanding its context and allowing a further adoption. For the representation of processes and agents, also the compatibility with Business Process Model and Notation (BPMN) [23] has been taken into account. In particular archives closely linked to media production institutions (e.g., broadcast archives, stock footage libraries) increasingly use service oriented architectures modelled using business processes.

While MP-AF has been defined in the context of MPEG, its use is not strongly linked to the use of other MPEG technologies for the content being preserved. For example, in the broadcasting domain, SMPTE MXF [24] has become the most widely used file container, and MP-AF can be used for representing preservation information of MXF files, whether the contained bitstreams are encoded using MPEG formats or not. Also, as described above, MP-AF allows for the inclusion of descriptive and technical metadata in different formats.

In the following paragraphs, a description of the relations to the three mentioned data models is reported, as displayed in Figure 4. In the figure, every element coming from other models is labelled with the corresponding prefix (i.e. premis, prov, ccdm) in order to disambiguate terms and avoid name mangling. Associations (connected lines with different symbols) are written following the UML2 [25] notation with the more general meaning (i.e. where it was not possible to define more precise relationships, a simple dependency with dashed lines has been adopted).

5.4.1 PREMIS

The compatibility of the MP-AF data model with the Object-Event-Agent structure in PREMIS is important in order to support organisations holding some amounts of audiovisual content, but which is not their main asset (e.g. National Libraries may preserve some audiovisual content, but their core assets are the book collections). Moreover the interoperability increases with the changes planned for the upcoming version 3 of PREMIS. As shown in Figure 4, the central element of the data model is the premis:IntellectualEntity that in MP-AF is the PreservationObject i.e. the entity that the model is describing with preservation metadata. In Figure 4 an UML dependency (dashed arrow) has been depicted connecting the two elements. Actually the PreservationObject is a child of the abstract element Item that in PREMIS can be considered as a child of premis:Object. The MP-AF Representation, File, Bitstream and UsageRights have quite straightforward PREMIS counterparts: the premis:Representation, premis:File, premis:Bitstream and premis:Rights. Concerning the latter, the MP-AF is more expressive because it can express usage rights (the rights expressed in premis:Rights) but can also express the ExploitationRights, i.e. much more complex rights (such as contracts) that can prevent many operations on the PreservationObject and must be captured as well.

The MP-AF Operator has the related element premis:Agent. In this case, MP-AF has decided to discriminate between human beings and machines, that is not directly possible in PREMIS. Hence the MP-AF Operator is a superclass of Agent for human beings and of Tools for software or other virtual actors. It follows that the premis:Agent had to be mapped to the more general parent class Operator. The MP-AF Activity, which is quite general, can be mapped to the premis:Event, that is associated to the premis:Agent performing or involved in the event as well as the Activity is associated to the Operator in MP-AF.

5.4.2 W3C PROV

The MP-AF data model is fully compatible with the Entity-Activity-Agent structure in the recently completed W3C Provenance data model [13]. The PROV data model is much smaller but actually the MP-AF element Item can be mapped to the prov:Entity. That means that all the MP-AF sub-elements of Item such as PreservationObject, Representation, Essence and Bitstream are mapped to prov:Entity as well. The prov:Activity performed on the prov:Entity can be mapped to the MP-AF Activity, with more or less the same semantic. The prov:Agent (as described in PREMIS) can be mapped to the MP-AF Operator. The MP-AF Group of several PreservationObject is represented in PROV as prov:Collection.

As with PREMIS, a direct mapping of MP-AF Asset does not exist, and only the more general PreservationObject can be mapped.

5.4.3 EBU CCDM

A mapping between the main entities in MP-AF and the EBU Class Conceptual Data Model (CCDM) has been established, aligning the model with a set of metadata models and formats implementing CCDM. The core element of CCDM is the ccdm:EditorialObject which is straight forwardly mapped to the MP-AF central element PreservationObject. CCDM, compared to the previous data models, is the closest to the audiovisual professional domains (EBU) and have more corresponding elements in the MP-AF model. For that reason we have immediate correspondence of MP-AF Asset (child of PreservationObject) to the ccdm:Asset, the Essence to the ccdm:Essence, the Group to the ccdm:Group, the Agent to the ccdm:Agent, the MediaLocator to the ccdm:Locator. Since Essence is a parent of the MP-AF Bitstream, it can also be considered mapped to the ccdm:Essence as well. Concerning the Rights, CCDM is able to represent complex rights, actually IPRights and the element ccdm:IPRights is closer to the MP-AF ExploitationRights rather than the simpler UsageRights. MP-AF Representation can considered the counterpart of ccdm:MediaResource.

5.4.4 Packaging MP-AF and Content

MP-AF does not specify a format for packaging metadata together with content, as different types of organisations preserving audiovisual content have different needs concerning the package format. The MP-AF implementation guidelines, which are currently being developed, describe how MP-AF can be embedded in some common container formats. Those include the MPEG Professional Archive Application Format (PA-AF), the Material Exchange Format (MXF [24]) and the Archive Exchange Format (currently under standardisation, to become SMPTE ST 2034).

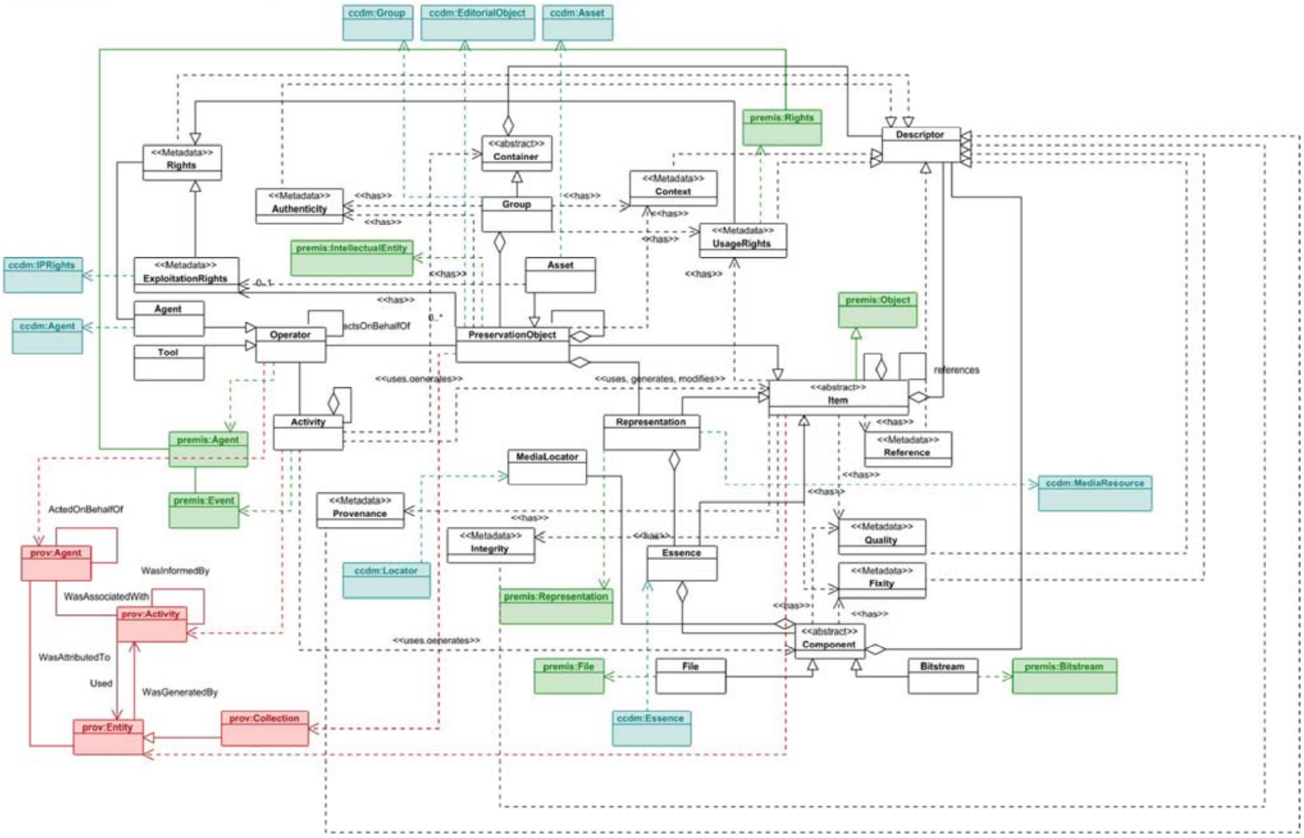


Figure 4: Visualisation of mapping of MP-AF entities to PREMIS, W3C PROV and CCDM.

6. CONCLUSION AND FUTURE WORK

This paper presents the work on a novel metadata MPEG standard specifying information needed for preservation of multimedia content, named Multimedia Preservation Application Format (MP-AF). The standardisation process is currently ongoing on committee level. The standard is expected to be completed in 2015 [26].

The presented model is able to fix the gaps already identified for representing preservation metadata of audiovisual contents and is sufficiently flexible to support various preservation workflows without imposing constraints on preservation environments or demanding changes of current models. It enables full interoperability between widely adopted preservation metadata schemas, various content structures and process models.

In next months the MPEG Multimedia Preservation group will evaluate the effectiveness of MP-AF in managing the multimedia preservation information in operative environments and use cases, in order to include samples and guidelines and allows an easy adoption in several different contexts.

7. ACKNOWLEDGMENTS

The authors would like to thank all the members of the MPEG Multimedia Preservation AHG. The presented work has been partially supported by the projects Presto4U (grant no. 600845) [4] and DAVID (grant no. 600827), both funded by the European Community under the 7th Framework Programme.

8. REFERENCES

[1] The Presto project – Preservation Technology for European Broadcast Archives, http://presto.joanneum.ac.at/ projects.asp. Last accessed 11 April 2014.

[2] The PrestoSpace project, An integrated solution for Audio-visual preservation and access, http://www.prestospace.org. Last accessed 11 April 2014.

[3] The PrestoPRIME project, http://www.prestoprime.eu. Last accessed 11 April 2014.

[4] Presto4U project – European Technology for Digital Audiovisual Media Preservation, http://www.prestocentre.org/4u. Last accessed 11 Apr. 2014.

[5] UNESCO International Conference: “Memory of the World in the Digital Age: Digitization and Preservation”, http://www.unesco.org/new/en/communication-and-information/resources/news-and-in-focus-articles/all-news/news/international_conference_memory_of_the_world_in_the_digital_age_digitization_and_preservation. Last accessed 11 April 2014.

[6] Moving Picture Expert Group - http://mpeg.chiariglione.org. Last accessed 11 April 2014.

[7] Consultative Committee for Space Data Systems. 2012. Reference Model for an Open Archival Information System (OAIS). Washington, DC: CCSDS Secretariat.

[8] Van Garderen, P. 2010. Archivematica: Using micro-services and open-source software to deliver a comprehensive digital curation solution, iPres2010 Conference.

[9] ISO/IEC 23000-6:2009, Information technology – Multimedia application format (MPEG-A) – Part 6: Professional archival application format.

[10] PREMIS Editorial Committee (2012). PREMIS Data Dictionary for Preservation Metadata, version 2.2, http://www.loc.gov/standards/premis/v2/premis-2-2.pdf.

[11] PREMIS Ontology, http://id.loc.gov/ontologies/premis.html. Last accessed 11 April 2014.

[12] Quality Control Test Definitions, EBU Tech 3363, Aug. 2013.

[13] Moreau, L. and Missier, P. (eds.). 2013. PROV-DM: The PROV Data Model, W3C Recommendation, http://www.w3.org/TR/2013/REC-prov-dm-20130430/

[14] CASPAR Consortium. 2008. Deliverable D2301 Report on OAIS-Access model, CASPAR-D2301-RP-0101-1_3, http://www.casparpreserves.eu/Members/cclrc/Deliverables/report-on-oais-access-model/at_download/file.pdf.

[15] Guercio, M., and Salza, S. 2013. Managing Authenticity through the Digital Resource Lifecycle. Digital Libraries and Archives, pp. 249-260.

[16] Bekaert, J., De Kooning, E., & Van de Sompel, H. 2006. Representing digital assets using MPEG-21 Digital Item Declaration. International Journal on Digital Libraries, 6(2), 159-173.

[17] Manepalli, S., Manepalli, G., and Nelson, M. L. 2006. D2D: Digital Archive to MPEG-21 DIDL, CoRR, http://arxiv.org/abs/cs/0602059

[18] PROV-O: The PROV Ontology, W3C Recommendation, April 2013, http://www.w3.org/TR/2013/REC-prov-o-20130430/

[19] Deliverable D2.1.3 AV Data Model. https://prestoprimews.ina.fr/public/deliverables/PP_WP2_D2.1.3_AV_Data_Model_R0_v1.00.pdf. Last accessed 11 April 2014.

[20] EBU Class Conceptual Data Model (CCDM), v1.0, Oct. 2012, http://tech.ebu.ch/docs/tech/tech3351.pdf

[21] ISO/IEC 15938-9:2005/Amd1:2012, Information technology – Multimedia content description interface (MPEG-7) – Part 9: Profiles and levels, AMENDMENT 1: Extensions to profiles and levels (Audio visual description profile (AVDP).

[22] ISO/IEC 21000-3:2003/Amd 2:2013, Information technology – Multimedia framework (MPEG-21) – Digital item semantic relationships.

[23] OMG. 2011. Business Process Model and Notation (BPMN), Version 2.0.

[24] SMPTE ST 377-1:2011, Material Exchange File (MXF) – File Format Specification

[25] UML. 2012. Unified modeling language, version (2.4.1), ISO/IEC 19505-1 and 19505-2.

[26] MPEG Multimedia Preservation Application Format, http://mpeg.chiariglione.org/standards/mpeg-a/multimedia-preservation-application-format. Last accessed 11 April 2014

Access and Preservation in the Cloud: Lessons from Operating Preservica Cloud Edition

Kevin O'Farrelly
Tessella
26 The Quadrant
Abingdon Science Park
Abingdon, UK
+44-1235-555511
Kevin.O'Farrelly@tessella.com

Alan Gairey
Tessella
26 The Quadrant
Abingdon Science Park
Abingdon, UK
+44-1235-555511
Alan.Gairey@tessella.com

James Carr
Tessella
26 The Quadrant
Abingdon Science Park
Abingdon, UK
+44-1235-555511
James.Carr@tessella.com

Maïté Braud
Tessella
26 The Quadrant
Abingdon Science Park
Abingdon, UK
+44-1235-555511
Maite.Braud@tessella.com

Robert Sharpe
Tessella
26 The Quadrant
Abingdon Science Park
Abingdon, UK
+44-1235-555511
Robert.Sharpe@tessella.com

ABSTRACT

The archival community has recently been offered a series of cloud solutions providing various forms of digital preservation. However, Perservica is unique in providing not just bit-level preservation but the full gamut of digital preservation services that, up until recently, were available only to organizations using a system installed on-site following on from a complex, and potentially risky, software development project. This "new paradigm" [1] thus offers a zero capital cost "pay as you go" model to perform not just bit-level preservation but also "active preservation" [2]. This short paper will describe the practical difficulties of providing and operating such a comprehensive service in the cloud.

A cloud system's advantage is to reduce the need for capital costs (since hardware and software are rented not bought up front) and system maintenance (since this is provided by the system's provider). To reduce costs further a system can share multiple organizations' content on a single operational instance. However, this instance must maintain each such tenant organization's isolation (i.e. one organization's content must not be exposed to any others). In addition each tenancy must be able to control its own processes without being able to compromise those of other tenants. This leads to the need for some degree of tenancy administration (without placing on each tenant a large burden of administration that is best handled at the system level).

The need to move bulk content across the internet as part of ingest cannot be avoided but the remaining ingest functionality can be performed either prior to upload (through a downloadable client-side tool) or server-side (through comprehensive workflows). Some ingest streams (e.g., web crawling) in fact can be considerably eased by using the cloud since an organization's local internet bandwidth is no longer relevant.

Other OAIS functional entities (preservation planning, data management, administration and storage) can all be performed without the need to move content across the internet. Access can be provided in a variety of forms including those suitable for archivists and those suitable for the general public. It is also possible to render content server-side to minimize the need for download.

Importantly, it is also possible to export an organization's entire content thereby providing a suitable "end of life" route to move to a different digital preservation system.

General Terms

Infrastructure, communities, strategic environment, preservation strategies and workflows, digital preservation marketplace, case studies and best practice.

Keywords

OAIS, Bit-level Preservation, Logical Preservation, Active Preservation, Cloud

1. INTRODUCTION

There has been a recent trend towards deploying and utilizing software systems in the cloud. In particular, digital archiving and preservation solutions are now available in the cloud. Cloud-based software systems (and digital archiving and preservation solutions in particular) have some distinct advantages and disadvantages over local deployment. This short paper compares and contrasts the experiences of developing solutions both on an organization's site and via a shared tenancy system in the cloud.

Note that in this paper, the term 'the cloud' is used to refer to public cloud instances, where services are made available over a publicly available network. While private clouds (i.e. cloud infrastructure operated solely for one organization) are similar to public clouds, many of the issues (legal, hardware provision and elasticity in particular) are different.

2. METHODOLOGY

In order to be able to discuss general issues that can occur with cloud systems and how it is possible to address these, it is necessary to have experience. This paper relies on Tessella's experience of developing and running both on-site and cloud-based preservation systems (Preservica). Hence, issues are discussed in general first and then (where appropriate) the Preservica solution to these issues is outlined.

Tessella's on-site preservation system (using the SDB software, recently rebranded as Preservica Enterprise) has been developed over about a decade and is deployed on-site by a number of leading archives and other memory institutions around the world. This allows bespoke functionality to be added to the system's core functionality in order to deliver a system that meets the specific, true needs of the organization.

The cloud-based Preservica service was launched in June 2012 and utilizes the same core software. It is deployed within Amazon Web Services cloud offerings.

3. CHOOSING THE CLOUD

There are a number of features that are important in determining whether or not to use the cloud for a digital preservation system.

3.1 Legal constraints

The use of a cloud solution means that content is stored away from an organization's own site. This may (or may not) be an issue depending on the nature of the content stored, the mandate of the organization and the legislative and regulatory framework in which they operate. The complex topic of intellectual property rights is covered in more details in other places [3].

The single biggest concern seems to be jurisdiction, with, for example, US institutions reluctant to let their content leave the United States and most European institutions reluctant to let their content leave the European Union. To get around this issue Preservica currently (March 2014) is deployed in two separate instances: one on the East Coast of the United States and the other in Dublin in Ireland.

Of course other organizations will have other constraints (e.g., defence contractors are unlikely to be willing to allow their information to be stored in a public cloud) that may prevent them from using the cloud.

3.2 Hardware & Elastic Computing

One of the advantages of cloud systems is that it is not necessary for an organization to purchase or maintain its own hardware. This removes the need for a capital budget and to have to make (often quite technical) purchasing decisions. It also removes the need to have to decide when it is necessary to perform a hardware upgrade (and to pay the capital cost associated with such an upgrade).

Cloud services are usually elastic. This means it is possible to add additional computational resources to expand computing capability. In the case of Preservica the core software works by passing the 'heavy loading' tasks to an array of job servers via a queuing system. This means that both on-site and cloud-based systems are known to scale very well. Of course such scaling comes at a cost whether it is via purchased, on-site hardware or rented, virtual servers in the cloud. One of the advantages of the cloud is that it is possible to rent servers for just the time that they are needed meaning that, for example, it is possible, to use the servers needed to process a backlog or a temporary ingest surge and then stop paying for them after that point.

In the case of buying cloud-based software-as-a-service each user is sharing processing resources with other users. Thus, it is the responsibility of the provider to ensure that sufficient resources are available to cope with steady loads and to deal reasonably with peak demands. Typically this will be monitored via a service level agreement (SLA) determining not just availability but also reliability whilst also specifying any limitations on, say, processing load that the tenants cannot break without sufficient prior agreement (to allow the service provider time to provision for it) and, potentially, payment.

3.3 Tenancies and Tenancy Isolation

Typically a cloud-based, software-as-a-service offering relies on economies of scale as hardware and administration costs are shared across all clients of the service. However, this means that clients of this service also share the same infrastructure, raising the potential for security breaches.

Hence, each organization utilizing the Preservica service becomes a 'tenant' within a selected instance. It is vital that these tenants remain isolated from each other and are not able to see each other's contents or to be able to tell the workflows etc. run by each other. Preservica has undergone extensive design reviews and a rigorous testing program to ensure tenant isolation.

3.4 Exit Strategy

Another very important aspect to consider in choosing a cloud system is how organizations will be able to move between providers. This is important since the cloud is still young and thus can be expected to evolve quickly. In order to be able to gain advantages from these changes, it is important that organizations don't become locked into arrangements that are very difficult to break for either contractual or technical reasons.

Preservica guards against this by allowing a full export of content with related metadata in a published AIP package format. This export process can be configured to allow alternative metadata schemas to be used and/or alternative packaging approaches. This allows great flexibility in how to export and thus in ability to import into a successor system.

3.5 Capital vs. Revenue Costs

Of course, a lot of decisions need to balance costs with the ideal functionality.

Typically, the cost of utilizing a standardized, full OAIS system via software-as-a-service in the cloud is much lower than the cost the more traditional alternative: owning and operating an on-site system, which has similar functionality but is highly configurable.

However, in certain circumstances it is possible for the economics to change in favor of the latter type of system, even considering the overheads involved. This is because the operational costs of a highly configured system can be greatly reduced and therefore overcome the two big overheads in setting up an on-site system: equipment capital costs and software capital costs. We will consider the hardware and software aspects of this in turn.

Generally the cost of renting cloud-based hardware is lower than the cost of buying and running an equivalent set of servers on site. However, at high storage volumes the economics of an organization running its own system begin to be comparable to, or even cheaper than, those of using a cloud-provided one. When taken together with the simplified exit strategy, this could lead to a decision to use an on-site solution. In addition, many organizations may have invested in on-site hardware already which, even if it needs upgrading, might still be cheaper than a compete transfer to the cloud.

Another potential overhead for an on-site solution is the capital cost needed to procure, develop and configure the system in the first place. Although a cloud, software-as-a-service system removes the need to pay these costs, by its very nature such a system must be generic. An on-site system, in contrast, can be built to meet an organization’s exact needs (ideally based off an existing, flexible starting system). For example, many of Tessella’s customers have procured systems to completely automate the process of ingesting very high volumes of materials using ingest workflows configured to work with the peculiarities of each source (e.g., to interpret the output of a digitization stream correctly and then ingest it). This can reduce the effort needed for ingest significantly and can produce a very high payback over the use of a more generic system that requires a large amount of intelligent user input in order to interpret the sources for each ingest of new material. Of course, having developed such software it need not necessarily be operated on-site and could be deployed in the cloud.

Hence, the decision on whether to use the cloud or not, is often a balance between one-off capital costs and on-going revenue costs. This involves balancing hardware procurement costs, hardware operational costs, software procurement costs and software-related operational costs. All of these depend not only on a fundamental appraisal of what would be best in a ‘green field’ development but also on what an organization might already own (e.g., if it already operates its own server rooms, the operational cost of adding a few extra servers might be very low). This could lead to a decision that the optimal solution is software-as-a-service in the cloud, a customized solution deployed on-site, or a customized solution deployed in the cloud.

4. STORAGE

Many people associate the cloud with storage. Indeed, a basic requirement of a digital preservation system is to offer bit-level preservation. Cloud-based digital preservation systems allow organizations to make use of the economies of scale offered by storing content using infrastructure beyond the means of most individual organizations. It also means that the operating and administration costs are similarly reduced.

In the case of Preservica, the S3 storage services offered by Amazon Web Services are used by default. These services create multiple copies in geographically separated places and perform their own integrity checking. This allows Amazon to claim 99.999999999% durability, which compares favourably to almost any in-house storage arrangement. However, organizations with a mandate to retain content in perpetuity are, naturally, wary of such claims (not least because even if it is accepted that the technical risk is extremely low there is a probability of the system ceasing to exist for other reasons). Indeed some cloud-based storage services have gone bankrupt and thus no longer exist.

To get around this issue, most cloud-based offerings allow organizations to choose to store copies in alternative storage systems. In Preservica’s case this can include the ability to hold a local copy using a ‘copy home’ storage mechanism (using ftp to write content back to hardware controlled by the host organization).

No system can offer a 100% guarantee. Hence, while it is tempting to continue to add more storage options, the ultimate goal will remain unachievable. Some providers do offer an insurance-backed guarantee. However, even here, it must be remembered that, as with other insurance, while a claim might lead to monetary compensation, this will not recover what has been lost, and it will still be necessary for an assessment of the value of what has been lost to be made prior to any claim being paid.

Ultimately, therefore, the appropriate storage policy is a compromise between costs and risks. Preservica allows this balance to be controlled differently based on appropriate criteria. Hence, a storage policy module allows organizations to choose different strategies for different content files (e.g., for digitization streams it might be appropriate to store the high-resolution master images in a cheaper storage system with low access capabilities, such as Amazon’s Glacier offering, while storing low-resolution, access copies in a highly available storage system such as Amazon S3).

Preservica has methods to allow content to be moved to allow for changes of policy, because of a change in the perception of risk, or to cope with a triggered risk (e.g., failure of a provider), or to optimize costs after a change in pricing. In the latter case it is important to weigh any costs of moving content (e.g., in bandwidth charges) against any potential savings.

5. ACCESS

Another important feature of most cloud solutions and digital preservation systems is access to content. The capabilities of systems vary here, but Preservica has two distinct offerings.

The first is an archivist’s user interface. This provides search and browse capabilities and offers a detailed view of the metadata of each entity (collections, records, files, and embedded objects within files) in the system. This includes the ability to view the audit trail and provenance of each entity. For records with multiple representations (e.g., those that have been migrated from one set of technologies to another) it is possible to compare the significant properties between each representation.

The second user interface is intended to be used by the general public to get live access to the parts of the collection they are allowed to see. This user interface deliberately only displays a subset of the available information about each entity (e.g., it excludes the audit trail) and only the representations intended for public consumption.

In addition, both user interfaces are capable of providing server-side rendering to allow users to view content without needing to download the full original file to their device. This is important in a cloud-based environment since downloads come at a cost and, depending on an individual’s internet connection, can be slow. It also allows complex technologies to be rendered (e.g., Preservica will render WARC files using the Wayback machine which otherwise would require a complex server setup to be used once the individual has downloaded such a set of files).

This approach of having two distinct user interfaces and therefore two very different user experiences is an example of the separation of concerns that is a feature of the cloud-based approach. It allows very different user communities to be supported from one system. The on-site approach to this issue has typically been to have separate systems (often from different suppliers) but this is harder in the cloud since the integration is much less efficient if systems are not co-located.

6. OTHER OAIS FUNCTIONAL ENTITIES

While most cloud-based systems just offer bit-level preservation and provide some form of ingest and access, these are only some of the functional entities in OAIS and are thus insufficient to meet its demands. Preservica provides a full OAIS solution in addition to Storage and Access described above. It has come about owing to the increasing maturity of the functionality of the core product. This ability to bring into the cloud functionality that was previously confined to on-site systems with a large bespoke element and significant capital costs, has been described as a "new paradigm" [1].

6.1 Ingest

A variety of routes are available including the ability to upload client-created SIPs (which can be created from ad-hoc content via a downloadable tool), create SIPs server-side from uploaded ZIP files and purely server-side ingest routes (e.g., web harvesting). All ingests pass through rigorous quality controls.

6.2 Data Management

This is highly flexible allowing users to describe the information using a schema of their choice and yet still search, view and edit the information [4]. In addition, it is possible to integrate with some external cataloguing systems.

6.3 Preservation Planning

This includes 'Active Preservation' [2] and includes the ability to perform both technical and conceptual characterization, determine which material is at risk either during ingest or at a later date, determine the most appropriate preservation plan, and then perform validated format migration at scale. This is controlled via a technical registry [5].

It is possible for users to download the output of migrations (in either production or test modes) should this be desired. However, it is important to note that, since validation is automated, this is not needed and thus, normally, migration does not require the content to leave the cloud servers,

6.4 Administration

If a cloud service is used it is not necessary for an organization to maintain its own technical administrative staff. This is especially valuable to smaller organizations since such tasks are often hard to resource. Even larger organizations find it hard to recruit, manage, train (and ultimately retain) technical staff such as database administrators. Sometimes such administration is out-sourced to a parent organization (e.g., a regional archive might rely on the central IT provision of the region's government). In these cases it can be hard for the needs of the smaller, client organization to be heard and understood by the administrators. Hence, for small and medium sized organizations, at least, there is a distinct advantage in buying a cloud-based service where the administration is performed by skilled and trained administrators who understand the needs of the system.

However, organizations still want (and need) to have some element of control. Hence, Preservica again separates the concerns and distinguishes system-level administration from tenant-level administration.

System-level administration involves managing availability, performing database backups, adding new patches and functionality etc. This is the responsibility of the service provider (Tessella in the case of Preservica Cloud).

The tenant-level administration (i.e. configuring functionality for an organization, determining which local metadata schemas to use etc.) needs to be controlled by the tenant and Preservica provides intuitive browser-based user interfaces to do so. This means that each organization can have control without having the burden of complex system administration.

7. CONCLUSION

This paper has presented some of the advantages and issues of running digital preservation services in the cloud. It shows that it is possible for this approach to offer a much-reduced entry barrier to organizations performing digital preservation without the need to compromise on demanding a full OAIS solution (i.e. both logical and bit-level preservation).

There are a number of technical challenges that have been overcome in the development of a cloud-based digital preservation service. They include:

- Enabling a carefully considered exit strategy.
- Allowing multiple storage options driven by an automatable storage policy.
- Allowing different access functionality for different classes of user, especially avoiding the need for download where possible.
- Providing full OAIS functionality on top of storage and access (i.e. not just bit-level preservation).
- Separating system-level administration (carried out by the supplier) from tenant-level administration (carried out by the tenant organization).

8. REFERENCES

[1] Adrian Brown. 2013 Practical Digital Preservation. Facet Publishing, London, UK.

[2] Sharpe R and Brown A. Active Preservation. Lecture Notes In Computer Science, 2009, Proceedings of the 13th European conference on Research and advanced technology for digital libraries, Corfu, Greece, Pages: 465-468.

[3] Andrew Charlesworth. 2012. Intellectual Property Rights for Digital Preservation. DPC Technology Watch Report 12-02

[4] Alan Gairey, Kevin O'Farrelly and Robert Sharpe. 2012. Towards seamless integration of Digital Archives with source systems. In *Proceedings of International Congress on Archiving* (Brisbane, Australia, 20-24 August 2012).

[5] Maïté Braud, James Carr, Kevin Leroux, Joe Rogers and Robert Sharpe. Linked Data Registry: A New Approach to Technical Registries. Submitted to iPres 2014.

Sustainability Assessments at the British Library: Formats, Frameworks, & Findings.

Maureen Pennock
The British Library
Boston Spa
West Yorkshire
+ 44 (0)1937 546302
Maureen.Pennock@bl.uk

Paul Wheatley
Paul Wheatley Consulting Limited
Leeds
West Yorkshire
@prwheatley
paulrobertwheatley@gmail.com

Peter May
The British Library
Euston Road
London
+44 (0)20 7412 7199
Peter.May@bl.uk

ABSTRACT

File format assessments have been the subject of much debate in and outside of the preservation community in the past decade. Recognizing the unique structural, operational, and collecting context of the British Library, the Library's digital preservation team recently initiated new format assessment work to deliver recommendations on which file formats will best enable the preservation of integral, authentic representations of British Library collection content over the long term. This paper describes the work carried out to review previous assessments, identify appropriate sustainability categories and newly assess formats accordingly.

We posit that the relatively 'fuzzy' nature of a file format requires a relatively open-ended assessment framework and a nuanced understanding of preservation risk that does not solely lie with 'all-or-nothing' format obsolescence. We review other work in this area and suggest that whilst previous format assessment work has addressed a range of subtly different aims, experience has since indicated that some of the criteria used - such as considering number of pages in a format specification as a measure of complexity - may be invalid. British Library assessments are made on documented points of principle, for example, an emphasis on evidence-based preservation risks and the avoidance of numerical scores leading to comparisons between formats, and these have formed the base upon which sustainability categories are defined. We present these categories, which help to identify preservation risks or other challenges in the management of digital collections, and provide an overview of initial assessments of three formats: TIFF, JP2, and PDF. We acknowledge however, that implementation of preservation requirements, e.g., the use of particular preservation-justified file formats, must be balanced against other business requirements, such as storage costs and access needs, and argue that transparency of this format assessment process is fundamental if the resulting recommendations are to be fully understood in the future.

General Terms

Preservation strategies and workflows, specialist content types, case studies and best practice

Keywords

British Library, file formats, sustainability, assessments, transparency, preservation master

1. INTRODUCTION

The British Library is increasingly a digital library. Our long term digital repository already holds over 11,500,000 items and more are added every day. With acquisition comes responsibility: we must preserve and make this content accessible for our future users - as a national library, this is at the heart of our mission. Yet preservation of digital content is not straightforward, requiring action and intervention throughout the lifecycle far earlier and more frequently than for our physical collection. The digital preservation team at the British Library is responsible for addressing this to ensure that despite the challenges, we are able to preserve our digital collections for the very long term.

The nature of the work carried out by the digital preservation team has changed since it was established in 2005. This is due in part to changes in leadership and organisational structure, but more significantly as a result of growth in our knowledge and changes in operational context. Furthermore, our digital library system has matured significantly in the past eight years, as has our understanding of key non- or semi-technical digital preservation needs in the Library. In 2013, the Library published a new digital preservation strategy that addressed these changes. The strategy identified four strategic priorities that must be addressed by 2016 [1]:

1. *Ensure our digital repository can store and preserve our collections for the long term* - enhancing its preservation capabilities and devising preservation plans for collections stored within;

2. *Manage the risks and challenges associated with digital preservation throughout the digital collection content lifecycle* - clearly defining our preservation requirements and implementing preservation risk management practices across the lifecycle;

3: *Embed digital sustainability as an organisational principle for digital library planning and development* - planning and budgeting for preservation and sustainability from the point of acquisition;

4: *Benefit from collaboration with other national and international institutions on digital preservation initiatives* - embarking on appropriate collaborative endeavours and achieving maximum return on investment in terms of time, effort and financial commitment.

These strategic priorities are addressed in a programme of work led by the digital preservation team that identifies and aligns eleven core workstreams with one or more priorities. Workstreams are highly interdependent; most are collaborative and require input from colleagues in other areas of the Library (e.g. curators, content owners, developers, architects, and processing staff), though a small number are driven and delivered by the digital preservation team alone.

The remainder of this paper is focused on the *File Format Assessment* workstream, which is delivered by the digital preservation team and aligns primarily with strategic priorities one and two. The workstream assesses file formats for long term preservation risks and identifies preferred formats for the preservation of collection content stored in our long term digital repository. It should be noted that the File Format Assessment work described in this paper is only one of several activities (including Policy Development and Collection Profiling) that provides input to preservation planning exercises. File format assessments should not be used in isolation to drive preservation decision making.

2. FILE FORMATS & LONG TERM PRESERVATION

Despite many years of global digital preservation research, experimentation and practice, fundamental questions about file formats and long term preservation remain under discussion. This section will attempt to assess work, thought and comment from the wider digital preservation community in order to inform a sensible and practical approach to assessing file formats and ultimately preserving digital collections.

2.1 What is a "File Format"?

A number of sources in the digital preservation sphere, for example the Global Digital Format Registry (GDFR) [2], have defined a file format as a representation of an information model, typically with an implied assumption that a file format is a method of structuring information in a sensible way for storage and ultimately retrieval and use. In the case of some file formats, such as TIFF[1], specifications have been created that do describe a reasonably sensible information model, as well as how it should be realised into an instance of the format. This concept has been identified and exploited for preservation purposes and is evident in the design and usage of the JHOVE tool, which compares a file against its respective file format specification and reports discrepancies.

More recently, some within the preservation community have observed that the software that is used to create instances of file formats also plays a role in defining what a file format is. Furthermore, a reference implementation of a viewer for a particular format could provide a different definition of the format itself. For example, Sheila Morrissey describes the "violations" of Adobe's specification for PDF that are tolerated by the Adobe Acrobat Reader [3]. Some of these were described in an appendix to the PDF specification, but were subsequently removed when PDF received ISO standardization. Morrissey states "...these notes, while helpful, beg the question as to what we are to consider authoritative with respect to PDF format instances: the specification, or the behavior of the Acrobat reader application."

An alternative definition proposed by Andy Jackson defines a file format as "a formal language defined for the purpose of persisting and transmitting the state of computer programmes" [4]. This position has been illustrated particularly well with extreme examples, such as that of the early binary office formats which effectively provided a dump of the application's internal data structures [5]. Rather than representing a cleanly structured information model, these formats were little more than a dump of application memory to enable faster loading and saving on sluggish 1990's-era PC hardware. The lack of a preservation community created format validation capability is hardly surprising in these cases.

Defining an appropriate scope for what we understand as a single file format is challenging. Many versions of a single format can exist, sometimes maintaining a degree of backward compatibility but sometimes involving wholesale redesigns over time (e.g., Office formats). Other formats allow embedding or attaching of yet other formats, leading to the possibility of veritable Pandora's Boxes of multi-format data waiting to be opened by reluctant preservationists.

Clearly the concept of a file format is difficult to tie down, and is perhaps most usefully considered as a somewhat amorphous entity. Assessment mechanisms (and indeed the preservation work they inform) will therefore need to take into account the somewhat imprecise nature of the main target of this work.

2.2 What is File Format Obsolescence? Does it Exist? And if so, to what Extent?

The digital preservation challenge was clearly identified and addressed by the Museums, Libraries and Archives (MLA) community in the latter part of the 1990s. A central theme that emerged from this early work was the danger of format obsolescence. This was characterised in a widely referenced piece by Jeff Rothenberg in Scientific American in 1995 where he stated "digital documents are evolving so rapidly that shifts in the forms of documents must inevitably arise. New forms do not necessarily subsume their predecessors or provide compatibility with previous formats" [6]. At the time, the IT market was emerging from an era characterised by a multitude of computer platforms, many of which had disappeared in a relatively short space of time. This was particularly evident in the home computing market. In this climate, the message that file formats were at risk of obsolescence unsurprisingly took hold. It can still be found today as a core part of many digital preservation training resources.

In the last few years a more sceptical view of file format obsolescence has emerged. David Rosenthal has made the case that format obsolescence simply doesn't exist, and references web-era work that provides some evidence to back up this position [7]. Evidence that makes a case for the format obsolescence lobby is harder to come by. Extreme examples sought and investigated by Chris Rusbridge were quite quickly solvable with help from colleagues and other expertise via the internet [8]. Rusbridge states "It's worth noting that a lot of the 'official' advice on obsolescence that you might find is useless. Various sites will classify formats as obsolete that are still perfectly easy to open and migrate from. Indeed, I suspect that there's no really helpful way to classify obsolescence (I tried and failed)".

Working on the assumption that data in the vast majority of file formats will be readable with some degree of effort does not take into account two crucial issues. Firstly, what is the degree of effort to enable rendering, and what does it mean for an organisation such as the British Library? Secondly, even if a file format is readable, is the resulting rendering, migration or indeed emulation, anything like an authentic reproduction of the original?

As a national memory institution, the British Library must ensure that collections are accessible for future generations. The term "institutionally obsolete" suggests a file format that may be accessible with further effort but will not run on a typical (or perhaps vanilla) computer platform provided by an institution [9]. In terms of the British Library this may relate to the platforms provided in our reading rooms or assumptions made about software available for those accessing our collections remotely[2]. Addressing this challenge may not be straightforward and has been taken into account in the assessment methodology on which this document focuses.

[1] Where interchange between software applications and the need to address the lack of an appropriate non-proprietary still image format was seen as a key aim in its conception.

[2] Increasingly, this means a web browser.

A number of studies have examined the impact of changing methods of rendering over time, where file formats may still be accessible, but with perhaps some degree of change in the results. These include the work of the Digitale Bewaring Project [10] and the Rendering Matters report which concludes that the “choice of rendering environment (software) used to open or “render” an office file invariably has an impact on the information presented through that rendering. When files are rendered in environments that differ from the original then they will often present altered information to the user. In some cases the information presented can differ from the original in ways that may be considered significant” [11]. The effects of the rendering process or environment on files (of particular formats) must be taken into account when considering the viability of a given preservation approach. What aspects of a digital collection item must be preserved and how can a given format support that?

In the necessarily conservative domain of digital preservation, it seems unwise to completely dismiss a concept such as format obsolescence on the evidence presented. However there are genuine and significant preservation risks beyond the black and white delineation of format survivability and they should be taken into account in the assessment methodology.

2.3 The Role of ‘Preservation Masters’

It is not uncommon for legal deposit legislation to stipulate that hard copy deposits must be the best available edition of a work[3]. The term ‘best’ is open to interpretation, though in Library contexts it is generally taken to mean content of the highest quality and most suitable for purpose. For example, archival-quality paper is preferred over low-grade paper, large size books are often preferred over small ones, complete versions are generally preferred over partial ones, and originals are preferred to copies[4]. Best editions are generally selected for their longevity and usability, both of which are important selection criteria for Libraries operating over the very long term.

‘Best’ editions remain significant in a digital environment. Digital content is liable to degrade in a similar fashion to hard copy, though in a shorter time frame, and although institutional obsolescence may not be imminent, it is inevitable eventually. The potential longevity of content is an essential consideration in institutions preserving for the long term. The same may be said of usability, where high-quality reproducibility and mutability, automated analysis, detailed searching and content enhancement all offer far more potential to the user than with physical copies. Our experience at the British Library is that in a digital environment, versions of collection items are often differentiated by format or format resolution, making format a key factor in determining best quality.

Preservation Masters play the role of our ‘best’ available digital editions at the British Library. The concept of a Preservation Master is not new, existing already for both physical and digital collections[5]. Preservation Masters are rich representations of a digital collection item with high levels of information content, which serve to meet both preservation needs and user needs by enabling the creation of derived files with minimum loss.

3. FORMAT ASSESSMENTS ELSEWHERE

File format assessments as a means to guide preservation activities have been ongoing in the preservation community since the latter part of the 1990s. They remain a hot topic in the community at the time of writing:

- The SCAPE Project presented a paper describing the File Format Metadata Aggregator (FFMA), an expert system to collate and assess file format information at iPRES2013 [12];
- The University of North Carolina is conducting research to gather expert opinion on file format risk;
- The 2014 National Agenda for Digital Stewardship identified "File Format Action Plan Development" as a specific priority “for infrastructure investment" [13];
- Lee Nilsson, the National Digital Stewardship Resident at the Library of Congress, recently provided an introduction to “File Format Action Plans”, which references much of the existing work in this area [14]. While not adding much new to the debate, it does indicate a commitment to follow up on the priority identified by the Library of Congress.

File format assessments have, however, been emerging for a number of years. Other notable work includes:

- The Florida Centre for Library Automation's File Format Background assessments and quite practically focused Action Plans that were developed from 2003 [15];
- The Library of Congress's widely referenced File Format Sustainability Factors [16];
- The National Library of Australia's AONS work, that attempted to score preservation worthiness [17]. The NLA subsequently moved away from this approach;
- Archivematica which realises file format migration on ingest (sometimes referred to as normalisation) based on a Format Policy Registry [18];
- Far less detailed file format guidance (albeit with obvious elements that can be traced back to the more comprehensive works referenced above) can be seen on innumerable sites across the web, for example the MIT Libraries Formats for Long-Term Access [19].

3.1 Theory versus Evidence

Johan van der Knijff notes that the criteria used in assessment approaches, such as that of the Library of Congress and the UK National Archives, “are largely based on theoretical considerations, without being backed up by any empirical data. As a result, their predictive value is largely unknown” [20]. Whilst such theoretical considerations may seem convincing, basing recommendations on real-world evidence provides a much more reassuring approach to preserving digital collections.

Where automated, top down approaches (such as the FFMA expert system) have the potential to replace expert analysis, there is considerable danger of poor, or possibly even catastrophic, preservation actions being taken. There are a number of documented (and anecdotally many more undocumented) examples of PDF migration implemented to ensure JHOVE provided a “valid and well formed” validation result for each preserved file, where there was little or no evidence of the need to

[3] See for example http://www.bl.uk/aboutus/legaldeposit/printedpubs/depositprintedpubs/deposit.html, and http://www.slsa.sa.gov.au/site/page.cfm?c=4702.

[4] United States Copyright Office *Best Edition of Published Copyrighted Works for the Collections of the Library of Congress*: http://www.copyright.gov/circs/circ07b.pdf.

[5] See for example the Preservation Policy of the National Library of Australia, 4th Edition: http://www.nla.gov.au/policy-and-planning/digital-preservation-policy.

take action given the tolerance of PDF viewers to many of the issues JHOVE identifies [21]. Given the potential for loss of important data when unnecessary format migration is applied (particularly given the woefully inadequate facilities for verifying the accuracy or quality of format migrations), this is particularly concerning. Van der Knijff notes alarm at "recurring attempts at reducing format-specific preservation risks to numerical risk factors, scores and indices"[20]. He goes on to provide an example from his own institution where a format assessment model [22] led to the adoption of JP2 instead of TIFF as the preservation format for digitised still image masters. A number of JP2 format risks were simply unknown at the time of the assessment and only became clear when the organisation worked with the format in practice. Van der Knijff summarises that "None of these problems were accounted for by the earlier risk assessment method (and I have a hard time seeing how they ever could be)!" This also lends support for an evidence backed approach, making recommendations based on empirical results; however, care should still be taken not to simply reduce such evidence to a numerical comparison between formats.

Archivematica is an example of a preservation system that implements file format normalisation on ingest to a repository. The Archivematica Format Policy Registry identifies which formats should be normalised, separately noting formats used for preservation and access [18]. They state that their "preservation formats are all open standards. Additionally, the choice of preservation format is based on community best practices, availability of open-source normalization tools, and an analysis of the significant characteristics for each media type". While the Registry usefully links to further detail and results from small scale testing, some of the normalization operations recommended are known to result in loss of fidelity, for example, transforming PDFs to PDF/A (which precludes some interactive content and hence would lead to data loss in files should normalization occur) or transforming GIF to TIFF (where the latter does not support the more unique animation properties of the former). The Registry justifies the PDF transformation by noting that "PDF/A is the only version of PDF recommended for long-term preservation". In a study of file format guidance from academic repositories in the US, Rimkus *et al* [23] reflect on the significant impact of particular sources of guidance, such as the frequently referenced and reused MIT Libraries Formats for Long-Term Access. They go on to state: "Comments made by repository managers during the data gathering period would imply that Archivematica is poised to play a similar role for the growing number of institutions that deploy it....Several digital preservation managers referred to Archivematica's ongoing file format policy registry and associated migration paths as the policies they intended to adopt at their own institutions".

Malcolm Todd's Digital Preservation Coalition Technology Watch Report: "File Formats for Preservation"[24] engages in a detailed discussion on the weighting and reconciliation of numerical scores for assessing formats based on a variety of assessment work. It concludes with support for score-based approaches, though the viability of these was later cast into doubt by Van der Knijff after practical experience with the approach at the Koninklijke Bibliotheek (see above).

There are a number of examples in which the application of assessment factors stop short of examining the practicalities of working with the format, some of which are listed above. Where this practical evidence is not available, proxies have been used without evidence that they are indeed linked to preservation risk - for example, a count of the number of pages in a file format's specification, or the number of applications that support a particular format. The former gives an impression as a crude measure of "format complexity", but arguably nothing more. Counting the huge number of pages in OOXML documentation might perhaps provide some indication of the sheer vastness of these formats but nonetheless it is woefully inadequate as a comparative measure between formats. The latter, on the other hand, can be simply misleading as many applications could rely on a small number of software libraries.

3.2 Clarity of Purpose and Audience

Format guidance to date has appeared to focus on addressing a range of subtly different aims, sometimes without clarity as to what those aims actually are. These include:

- Guidance that records the level of support that will be provided to data preserved within a particular repository (typically ranging from some kind of guarantee or best effort, through to bit preservation only);
- Guidance that targets contributors to digital repositories, sometimes recommending formats in which particular types of data should be submitted;
- Guidance that targets data creators, recommending formats in which particular types of data should be created;
- Justification and guidance for repository/preservation managers in implementing recommendations, possibly addressing format migration or normalisation.

Where these aims are unclear or, perhaps even more significantly, the target audience of the guidance is unclear, the potential for misuse becomes real. This becomes especially concerning where guidance is re-used outside of its original context, such as by another organisation. As the examples in the previous section indicate, file format assessments and resulting guidance can have a significant impact within the wider community, leading to the possibility of mis-informed preservation choices.

3.3 Misleading Measures

Adoption rates and (self)-documentation are common features in the assessment frameworks mentioned above that can be misleading if not properly understood.

A reference to the availability of documentation can be found in most of the existing file format assessment work. In the UK National Archives' "Selecting File Formats for Long-Term Preservation" Adrian Brown states that the "availability of format documentation is not, in itself, sufficient; documentation must also be comprehensive, accurate and comprehensible. Specifically, it should be of sufficient quality to allow interpretation of objects in the format, either by a human user or through the development of new access software" [25]. Brown also suggests that a "detailed judgment of documentation quality will require evaluation of the documentation itself". However the only way to be sure that documentation is sufficiently complete to enable development of new access software would be to develop and test new access software from it. This is a costly approach. Documentation is undoubtedly beneficial to have in some circumstances, but assessing or rating the quality of documentation is clearly problematic and so use in assessing the sustainability of file formats requires careful consideration. As van der Knijff states: "A problem with errors and ambiguities in format specifications is that they can be incredibly easy to overlook, and you may only become aware of them after discovering that different software products interpret the specifications in slightly different ways" [26].

The value of self-documentation (where sufficient metadata is present to aid in understanding and/or use of the format, without the need for additional attached metadata) is debatable for collections that reside within a modern digital repository with comprehensive support for attached metadata. While embedded metadata may provide some use in the event of catastrophic repository damage that might physically separate collections from their metadata, this is an eventuality that repository design, replication and backups aim to avoid. Conversely, where metadata is both embedded in a file and associated or attached in a repository, should it be kept consistent? To do so may require frequent modification to the collection object - a course of action in itself that introduces preservation risk, and hence is probably undesirable. If embedded and attached metadata is inconsistent its value becomes questionable. It therefore seems sensible not to take self-documentation into account in a format assessment of this kind.

Measuring "adoption" of a format in the wider world is clearly a difficult task. What level of adoption is sufficient? How might it be quantified? Observations about formats residing in niches, perhaps in conjunction with the availability or quality of software to render the format in question, could provide useful insight. The adoption of the JP2 format within the library community provides some interesting observations. At a digital preservation meeting at the Wellcome Library focusing on JP2 in 2010, comments from members of the audience suggested that a number of libraries within Europe had adopted JP2 "because that was what the British Library had done". It should be noted that the BL adopted JP2 for use in very specific high volume collections and otherwise still utilises TIFF. This example worryingly highlights the impact of hearsay and reputation over analysis and evidence. It also poses questions about analysis that might be based on generalised assessments of adoption. Despite growing numbers of MLA organizations adopting JP2 for storing digitized images (noting that the picture is somewhat muddied by JP2's attractiveness in not only reducing storage volume but also in potentially delivering content to remote users, thereby seeing some use as a preservation format, some as an access format and in some cases both), there remain serious concerns about the quality and sustainability of creation and access software [27]. Clearly measures of adoption in isolation can be misleading. Turning an impression of adoption into a numerical rating to facilitate relative scoring of formats could prove to be a dangerous approach. Approaches that draw conclusions based on surveys of existing advice should also be viewed with caution.

4. BRITISH LIBRARY FORMAT ASSESSMENT POINTS OF PRINCIPLE

Discussion around the issues above has been distilled into the following points of principle that inform the implementation of format assessments:

1. Clearly state the aims of the assessment, the target of resulting guidance and the circumstances within which guidance should be acted upon;
2. Be aware of the potential for file format obsolescence but proceed on the basis that catastrophic loss of access to a particular format will not usually be the most pressing preservation risk;
3. Published guidelines, policies and assessments have a ripple effect and are often reused without consideration of the underlying evidence or the influence of unique organisational requirements. Meta assessments that make recommendations based on surveys of what other organisations do, add a further level of obfuscation. Approach with caution

For assessments:

4. Focus on evidence-based preservation risks (for example, non-embedded fonts in PDF);
5. Focus on implications of institutional obsolescence which lead to issues maintaining the content over time;
6. Any recommendations to choose a preservation format different to the format in which the data was received must be backed up by strong empirical evidence of the benefits and risks involved;
7. Avoid assessment based on theoretical factors and avoid format-to-format comparisons using summarised sustainability factors (in particular numerical scoring based approaches).

On specific sustainability factors:

8. Measures of "documentation completeness" or quality are largely meaningless and should be avoided;
9. Self-documentation should not be considered as an assessment factor. Documentation availability should be considered with a view to supporting likely preservation processes rather than as a judgment of preservation worthiness.

Many other organisations have exactly the same challenges in a different context. Assessments are therefore undertaken in an open and collaborative manner in order to increase the effectiveness of the decision making (based on greater contribution from an array of expertise) and minimise the resources required from the British Library.

5. SUSTAINABILITY CATEGORIES

The British Library assessment of file formats against sustainability categories identifies areas for concern rather than rating a format on a comparative scale. Practical guidance on mitigation practices for areas of concern is provided at the end of each assessment, though it should be noted that the capability (e.g., appropriate software tools) will not always exist to address all areas of concern. In some cases it is necessary to identify instead areas for experimentation with software tools and their impact on sample collections.

In summary, each file format assessment aims to provide evidence-based recommendations around use of a specific format, including whether or not a format is suitable as a Preservation Master within the British Library. Risks of using the format are identified and initial mitigation advice listed. Where there is uncertainty, this is clearly stated.

Sustainability categories considered in the assessments are as follows:

Development Status: An overview of the history, ownership, and current status of the file format.

Adoption and Usage: An impression of how widely the file format is used, with reference to usage in other memory institutions and their practical experiences of working with the format.

Software Support: *Rendering Software Support* - an overall impression of software support for rendering the format with reference to a) typical desktop software and b) current support on British Library reading room PCs; *Preservation Software Support* - an impression of the availability and effectiveness of software for managing and preserving instances of the file format, including a) Format Identification, b) Validation and Detecting Preservation Risks, c) Conformance Checking, d) Metadata Extraction, and e) Migration.

Documentation and Guidance: An indication of the availability of practical documentation or guidance with specific reference to the facilitation of any recommended actions

Complexity: An impression of the complexity of the format with respect to the impact this is likely to have on the organisation managing or working with content in this format. What level of expertise in the format is required to have confidence in management and preservation?

Embedded or Attached Content: The potential for embedding or attaching files of similar or different formats, and the likely implications of this.

External Dependencies: An indication of the possibility of content external to an instance of the file format that is complimentary or even essential to the intellectual content of the instance.

Legal Issues: Legal impediments to the use, management or preservation of instances of the file format.

Technical Protection Mechanisms: Encryption, Digital Rights Management and any other technical mechanisms that might restrict usage, management or preservation of instances of the file format.

Other Preservation Risks: Other evidence based preservation risks, noting that many known preservation risks are format specific and do not easily fit under any of the sustainability categories above.

Categories were defined prior to assessment and without consideration of any specific formats, in order to deliver a 'vanilla' set with no specific format bias. The detail of each category has been elaborated upon as a result of our experience in the initial assessments, but none have been deleted.

6. RESULTS

Six formats have been assessed to date: TIFF, JP2, PDF (including PDF/A), NTF (Ordnance Survey), JATS and ePub. Assessments typically take between 4 – 6 working days to complete, including background research. Results are issued in the first instance in the form of a report, which is subsequently condensed into a summary table for clarity and ease of dissemination. Due to space restrictions in this paper it is not possible to include more than summary discussions for the first 3 formats assessed. The full reports will be published elsewhere by the British Library in due course.

The TIFF assessment concluded that TIFF remains reasonably well suited to the simple task of the storage of digitised preservation masters, despite lacking many new bitmap file format features that have developed to support advances in graphics applications since the last significant changes to the format. Although there are preservation concerns with less well supported features that were introduced in version 6, baseline tags are well supported by software and well tested by many users both within and beyond the MLA sector. Implementation of a TIFF parser/profile conformance checker of a similar form to Jpylyzer [28] would be useful in performing assessments of trial runs in new digitisation projects and allow automated checking of subsequent production runs to the same standards. Detection of poorly supported TIFF extensions would also enable identification of problem content in deposited collections. Further investigation and/or collaboration with institutions interested in developing a "TIFFylyzer" and developers of the Kost-val validation application [29] should be explored.

JP2 fared less favourably than TIFF as a format for digitised preservation masters. Based on the evidence collected, the assessment concluded that JP2 is undesirable from a purely preservation-oriented perspective. JP2 is a niche format that has failed to see widespread adoption. As a consequence there is poor tool support and significant numbers of issues have been reported, despite the low rate of adoption. Obvious bugs in both the format and in software were not fixed before the preservation community adopted JP2 [30]. It is hoped that growing use by memory organisations and associated experience in working with JP2 will eventually lead to mitigation of most issues, but other problems may remain. In the meantime, if the benefits of JP2 (compression and delivery) are sufficient that it remains a desirable solution for storing digitised preservation masters, use of the format must be considered a significant risk. Ideally, mitigation of this risk requires investment in tools such as OpenJPEG to address the tool support concerns, and very thorough checking of all files in production settings. Mitigating JP2 preservation concerns comes with an associated cost and this should be taken into consideration in preservation planning activities where storage cost savings are likely to be significant.

PDF is a ubiquitous format in the contemporary computing world but widespread adoption, usage and software support has not led to the universal mitigation of preservation risks associated with this format. PDF files are frequently found to be invalid or badly formed and whilst the tolerance of most PDF rendering applications makes the impact of this situation difficult to measure, it should nonetheless raise a red-flag for preservation over the long term. A number of the other identified PDF risks have the potential to be catastrophic from a preservation point of view (such as encryption or missing font information, which could prevent access to content altogether). Strengthening our ability to detect these risks and ultimately developing trusted (and verifiable) means of fixing these issues in PDF files will be essential. That said, the severity and frequency of the risks identified in the full report remain relatively poorly understood. Existing published research has only begun to scratch the surface in revealing how these risks may affect an archive collection of PDF files (or not, as the case may be!). Research to apply validation tools to collections in order to more clearly identify genuinely problematic PDFs, or indeed discount identified risks whose frequency or impact is not significant, would help considerably to inform handling guidelines and potentially avoid overly prescriptive and potentially costly PDF fixing that has been adopted by some organisations. Testing of this sort is expected to take place over the course of 2014/15 in a Tool Assessment workstream, using collections identified as part of a Collection Profiling exercise (the subject of another paper submitted to iPRES 2014). The nature of the restrictions in PDF/A preclude preservation of some functionality and therefore its application will not necessarily suit every use case. For example, wholesale migration of a PDF collection to one of the PDF/A versions is unwise as functionality such as audio and video will be discarded. However, receipt of deposit of a PDF/A-1 may not raise significant preservation concerns as the PDF/A restrictions prohibit functionality associated with the preservation risks identified in the assessment - assuming of course that the PDF/A-1 files do indeed conform to the restrictions described in the PDF/A-1 standard. This is nonetheless a potentially dangerous assumption and one that may be difficult to test given concerns about PDF/A validation.

7. CONCLUSIONS

It is clear from reviews of earlier work that proxy measures of preservation risks are insufficient to capture the subtleties involved in practical digital collection management and long term preservation. Format assessments should be informed by thorough practical considerations and, insofar as is possible with long term investigations without a crystal ball, empirical evidence. This will only be possible at scale in a global community if we share not only our findings but also our aims, our context, and our underlying data. Otherwise we are doomed to repeat our failings. The conclusions of this work concerning the JP2 format are, we hope, an alarm bell for institutions choosing to preserve in this format primarily on the basis that others are doing so. Preservation Masters are the files from which future iterations of a digital collection item will be generated, and it is essential that their selection is fully informed.

Considering the applicability of the assessments to date to a much bigger and heterogeneous digital collection, as is the case at the British Library, it is further noted that assessments based around file formats *alone* reveal only some of the critical preservation issues that need to be addressed. Many digital collection items are compound in nature and may consist of a number of files, each possibly of a different format. Consideration must be given to all formats, their inter-relationships, and the compound object, for an assessment to be valid. The potential for a format to store different types of content must also be accounted for, as formats for digitised still images may likely have different requirements to formats for digitised manuscripts or born-digital images. Assessments of this sort are, however, the first step along that road and remain essential for memory institutions to understand why a given format is preferred over another, particularly those institutions with a mandate to preserve for the very long term. Transparency of the process is key to that understanding.

Finally, we observe the importance of the action taken as a result of an assessment. This work suggests a new and more nuanced approach is necessary to avoid the comparative scoring of format against format and the focus on format obsolescence without consideration for more subtle and pressing preservation risks. Assessments can provide an invaluable steer to essential preservation activities. This could take the form of specific handling guidance to mitigate clearly identified preservation risks, identification of preferred deposit formats for different types of content, further research and practical testing to fill gaps in existing understanding, or engagement with the responsible owner of a format to provide feedback on file format specification errors or ambiguities.

Ultimately a Preservation Master, with respect to a particular collection, can only be established through an effective preservation planning activity in which file format assessments provide only one of many essential information inputs.

8. ACKNOWLEDGMENTS

Thanks go to Johan van der Knijff of the National Library of the Netherlands for his tireless investigative work on which many aspects of this format assessment work were based. Thanks also go to Sheila Morrissey of Portico for her PDF expertise.

9. REFERENCES

[1] Pennock, M. 2012. *British Library Digital Preservation Strategy, 2013 – 2016*. British Library (May 2012). URL= http://www.bl.uk/aboutus/stratpolprog/collectioncare/discovermore/digitalpreservation/strategy/BL_DigitalPreservationStrategy_2013-16-external.pdf.

[2] The GDFR Ontology defined a format as "a byte-serialized encoding of an information model". The document is no longer available but is referenced here: http://www.digitalpreservation.gov/formats/intro/format_eval_rel.shtml.

[3] Morrissey, S. M. 2012. The Network is the Format: PDF and the Long-term Use of Digital Content. In *Proceedings of IS&T Conference Archiving 2012* (Copenhagen, Denmark, June 12 - 15, 2012). 200-203. Online ISSN: 2168-3204. URL= http://www.ingentaconnect.com/content/ist/ac/2012/00002012/00000001/art00044.

[4] Jackson, A. N. 2012. Tweet. URL= https://twitter.com/anjacks0n/status/167279401057255425.

[5] Spolsky, J. 2008. Why are the Microsoft Office file formats so complicated? (And some workarounds). *Joel On Software Blog* (2000 – 2014). URL= http://www.joelonsoftware.com/items/2008/02/19.html.

[6] Rothenberg, J. 1995. Ensuring the Longevity of Digital Documents. *Scientific American*. 272, 1 (1995). ISSN: 0036-8733.

[7] Rosenthal, D. 2012. Formats through time. *DSHR Blog* (2007 – 2014). URL= http://blog.dshr.org/2012/10/formats-through-time.html.

[8] Rusbridge, C. 2012. The PowerPoint 4.0 adventure: what did I learn? *Unsustainable Ideas Blog (2011 – 2013)*. URL= http://unsustainableideas.wordpress.com/2012/10/15/ppt-4-adventure-learning/.

[9] De Vorsey, K., and McKinney, P. 2010. Digital Preservation in Capable Hands. *Information Standards Quarterly*. 22, 2 (Spring 2010). ISSN 1041-0031. URL= http://www.niso.org/apps/group_public/download.php/4242/IP_DeVorsey_McKinney__Risk_Assessment_isqv22no2.pdf.

[10] Testbed Digitale Bewaring. 2010. *Migration: Context and Current Status*. White Paper. Digital Preservation Testbed project. (The Hague, December 5 2001). URL= http://en.nationaalarchief.nl/kennisbank/migration-context-and-current-status-2001.

[11] Cochrane, E. 2012. *Rendering Matters*. Technical Report. Archives New Zealand (Wellington, January 2012). URL= http://archives.govt.nz/rendering-matters-report-results-research-digital-object-rendering.

[12] Graf, R., and Gordea, S. 2013. A Risk Analysis of File Formats for Preservation Planning. In *Proceedings of the iPres 2013 Conference (*Lisbon, Portugal, September 2 – 6, 2013). PURL= http://purl.pt/24107.

[13] National Digital Stewardship Alliance (NDSA). 2013. *National Agenda for Digital Stewardship*. Report. NDSA (July 2013). URL= http://www.digitalpreservation.gov/ndsa/documents/2014NationalAgenda.pdf.

[14] Nilsson, L. 2014. File Format Action Plans in Theory & Practice. *The Signal Blog* from the Library of Congress (2011 – 2014). URL= http://blogs.loc.gov/digitalpreservation/2014/01/file-format-action-plans-in-theory-and-practice/.

[15] Florida Virtual Campus. 2012. *FCLA File Format Assessments*. Report from State University Library Services, Florida Virtual Campus. URL= http://fclaweb.fcla.edu/node/795.

[16] Library of Congress. *Sustainability of Digital Formats: Planning for Library of Congress Collections*. Digital Formats website. URL= http://www.digitalpreservation.gov/formats/sustain/sustain.shtml.

[17] Pearson, D., and Webb, C. 2008. Defining File Format Obsolescence: A Risky Journey. *International Journal of Digital Curation*. 3, 1 (July 2008). ISSN: 1746-8256. DOI = http://dx.doi.org/10.2218/ijdc.v3i1.44.

[18] Archivematica. 2014. Format Policies. *Archivematica wiki*. URL= https://www.archivematica.org/wiki/Format_policies.

[19] MIT Libraries. 2014. Data Management and Publishing. *MIT Libraries website*. URL= https://libraries.mit.edu/guides/subjects/data-management/formats.html.

[20] Van der Knijff, J. 2013. Assessing file format risks: searching for Bigfoot? *Open Planets Foundation Blog* (2010 - 2014). URL= http://www.openplanetsfoundation.org/blogs/2013-09-30-assessing-file-format-risks-searching-bigfoot.

[21] Leibniz Information Centre for Economics. 2013. Hunger for Automation – The first migration actions in our Rosetta Digital Archive (poster). *International Digital Curation Conference 2013* (Amsterdam, The Netherlands, January 14 – 17, 2013). URL= http://www.dcc.ac.uk/sites/default/files/documents/idcc13posters/Poster213.pdf.

[22] Rog, J. and van Wijk, C. 2008. *Evaluating File Formats for Long-term Preservation*. Technical Report. Koninklijke Bibliotheek, (Den Haag, The Netherlands, 2008). URL= http://www.kb.nl/sites/default/files/docs/KB_file_format_evaluation_method_27022008.pdf.

[23] Rimkus, K. et al. 2014. Digital Preservation File Format Policies of ARL Member Libraries: An Analysis. *D-Lib Magazine*, 20, 3⁄4, (March/April 2014). DOI= http://dx.doi.org/10.1045/march2014-rimkus.

[24] Todd, M. 2009. *File Formats for Preservation: A DPC Technology Watch*. Technical Report. Digital Preservation Coalition (London, October 2009). URL= http://www.dpconline.org/component/docman/doc_download/375-file-formats-for-preservation.

[25] Brown, A. 2008. *Selecting File Formats for Preservation*. Technical Report. The National Archives (London, August 2008). URL= http://www.nationalarchives.gov.uk/documents/selecting-file-formats.pdf.

[26] Van der Knijff, J. 2010. Ensuring the suitability of JPEG 2000 for Preservation. *Wellcome Library Blog* (2010 - 2011). URL= http://jpeg2000wellcomelibrary.blogspot.co.uk/2010/12/guest-post-ensuring-suitability-of-jpeg.html.

[27] Open Planets Foundation. 2013. Lack of Performant Open Source Decoding Software for JP2. *Open Planets Wiki* (2010 – 2014). URL= http://wiki.opf-labs.org/display/TR/Lack+of+performant+open+source+decoding+software.

[28] Open Planets Foundation and Koninklijke Bibliotheek. 2014. Jpylyzer. *JP2 validator and feature extractor website*. URL= http://openplanets.github.io/jpylyzer/.

[29] Röthlisberger, C. 2014. Kost-Val entry in the COPTR directory. *Community-Owned digital Preservation Tool Registry (COPTR)*. URL= http://coptr.digipres.org/KOST-Val.

[30] Van der Knijff, J. 2013. ICC profiles and resolution in JP2: update on 2011 D-Lib paper. *Open Planets Foundation Blog* (2010 – 2014).URL= http://www.openplanetsfoundation.org/blogs/2013-07-01-icc-profiles-and-resolution-jp2-update-2011-d-lib-paper.

Converting WordStar to HTML4

Jay Gattuso
Jay.Gattuso@dia.govt.nz
National Library of New Zealand
Cnr Molesworth & Aitken St
Wellington
New Zealand

Peter McKinney
peter.mckinney@dia.govt.nz
National Library of New Zealand
Cnr Molesworth & Aitken St
Wellington
New Zealand

ABSTRACT

In this paper, we describe the processing and verification work undertaken to migrate WordStar for MSDOS to HTML4 formatted files.

General Terms

Preservation strategies and workflows, specialist content types, digital preservation marketplace, theory of digital preservation, case studies and best practice.

Keywords

Preservation planning, preservation action, WordStar, HTML, significant properties, acceptable change, converters.

1. INTRODUCTION

In 2012, the National Library of New Zealand's (NLNZ) digital preservation business unit made the decision to undertake its first "in-anger" preservation planning and action activities to mitigate the risk to content that forms part of the Library's collections.

The criteria for the set was defined as (1) the format type should display a significant risk to its future use (2) of a manageable size; (3) from one collection group; and, (4) of a simple construction, (i.e. no compression or complex wrappers/containers).

In assessing the 137 uniquely identifiable formats[1] that currently appear in the digital preservation system, the best fit format for this work was WordStar. NLNZ holds 37 files that have been identified as WordStar for MSDOS formatted files.

We don't know exactly what version of WordStar files these are, as there are no signature based format identifications available for these file types. To that end, searching the system for all files with a .ws file extension would have resulted in the same corpus being constructed.

WordStar for MSDOS would be described by NLNZ as "functionally obsolete", meaning that it is highly unlikely that a normal user would have the tools to open and accurately render the content in a way that was in keeping with its original layout and intent.[2]

They are also all part of the same collection and therefore meet all of the criteria.

We had a second order requirement, to explore what a migration process feels like to all involved parties (technical, curatorial and managerial). This would serve as the starting point for more related activities in the future, and as such we wanted to ensure that we at least understand the basic framework that would underpin future migration work.

2. INITIAL ANALYSIS OF THE CONTENT

It was relatively simple to search the preservation repository for all the WordStar files. Simply searching the ~10 million files in the preservation repository for any of the PRONOM PUIDs that are registered against WordStar formats resulted in 37 files being identified. [3]

These files were retrieved from the system, and inspected in detail to ascertain their composition. The inspection demonstrated the following:

- there is no complex formatting or layout (e.g. tables), just text;
- "normally" encoded UTF-8 text is visible amongst the format structure;
- all the content is of a similar form, and is relatively straightforward to process;
- all files open OK with a not-quite contemporary copy of WordStar for MSDOS (see following paragraph);
- all files are transcripts of audio interviews that belong to our Oral History Unit. These files are highly restricted and cannot be shared outside of the Library at this time[4].

We also managed to locate a computer of approximately the correct age for the corpus. This was part of a relatively unmanaged collection of ICT equipment that has been put aside for testing purposes by the library.

[1] Identified using the DROID file format tools over 5 years.

[2] For a discussion on the Library's view on this, see [4].

[3] PRONOM PUIDs x-fmt/370, x-fmt/260, x-fmt/205, x-fmt/236, x-fmt/237, x-fmt/261, x-fmt/206 and x-fmt/262. [3].

[4] This by itself causes us problems for verifying our work. We cannot share these files for peer review, and any effort to create new sharable files may not result in the same "version" of WordStar files being created. This also precludes the ability to use any online conversion services to test their capability.

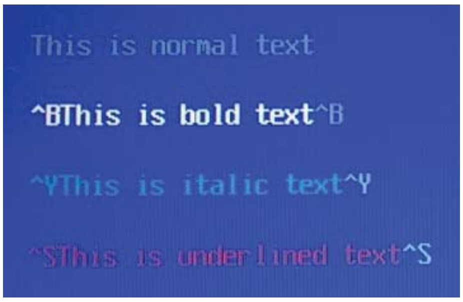
The computer is Toshiba Satellite T2130CT (circa 1995, running a 486 intel chipset, and MS Windows v3.11 / MSDOS v6).

We also found a copy of WordStar for MSDOS v5 on ebay.com, which was installed on the machine for testing / reference however, we do not know if any of the WordStar in the corpus were created with this version of WordStar for MSDOS.

WordStar, not unusually for word processors of its era (circa 1986), used a method of displaying markup on screen as tags, not unlike un-rendered HTML, rather than affecting the formatting visually as is more common today (rendered HTML). This means if an author wanted to bold format the words “`I am bold`”, it would appear on screen as “`^BI am bold^B`” rather than the formatting being applied directly to the text - “**`I am bold`**”. Of course, when the page is printed, the markup tags are not printed, but the intent of the formatting tags are.

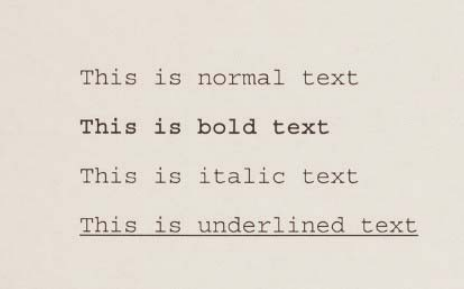
Text formatting is not the only marked up feature that can be found in the WordStar formatted content.

Some example files were created on the original hardware, printed, and imaged by our internal image services to allow us to demonstrate the visible difference between the screen view, and the printed page:-



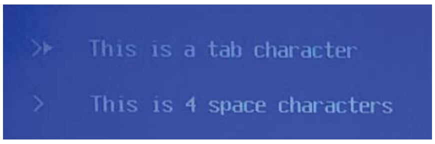
```
This is normal text

^BThis is bold text^B

^YThis is italic text^Y

^SThis is underlined text^S
```

Figure 1 - Screen shot: WordStar on screen markup



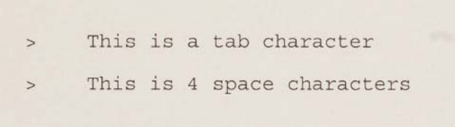
```
This is normal text

This is bold text

This is italic text

This is underlined text
```

Figure 2 - WordStar printed output



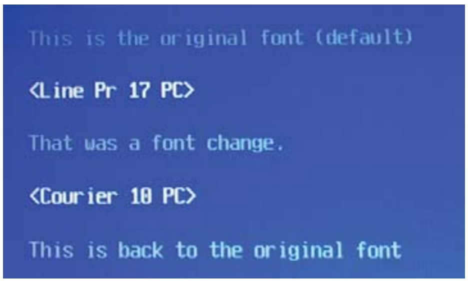
```
>▸   This is a tab character

>    This is 4 space characters
```

Figure 3 - Screen shot: WordStar on screen markup



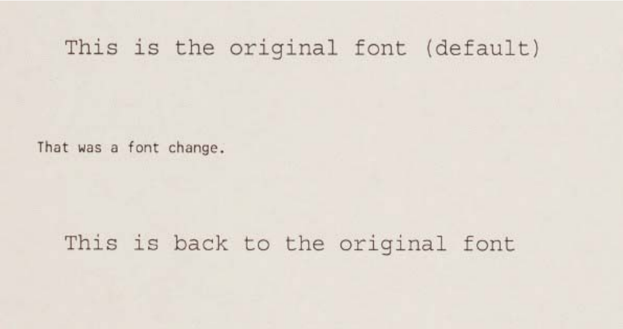
```
>    This is a tab character

>    This is 4 space characters
```

Figure 4 - WordStar printed output



```
This is the original font (default)

<Line Pr 17 PC>

That was a font change.

<Courier 10 PC>

This is back to the original font
```

Figure 5 - Screen shot: WordStar on screen markup



```
This is the original font (default)


That was a font change.


This is back to the original font
```

Figure 6 - WordStar printed output

The WordStar files in the corpus were inspected, and the following bytes were found, used in some way by WordStar to convey formatting or other information (see Table 1).

These all appear in the “control word” section of the UTF-8 text encoding standard.[6]

This was achieved by parsing the files in the corpus byte by byte, and returning any bytes that fall outside of the range of UTF-8 code points that have a normally associated printable glyph (`\x20` to `\x7e`)[5].

All of these code points needed to be addressed in some way to ensure that their meaning or purpose is properly conveyed by any converted files where applicable.

It was clear from the tested files opened with a “normal”/ modern text viewer, and the reference version of WordStar for MSDOS, that the control code-points in UTF-8 have an entirely different function in WordStar for MSDOS files.

Their individual functions in the WordStar files are compared to UTF-8 in the table below:-

[5] For the duration of this paper. any hexadecimal byte is represented by the hexadecimal value, preceded by “\x”.

Table 1 - Control byte comparison of UTF-8 and WordStar

Bytes	UTF-8	WordStar[6]
\x02	Start of Text [STX]	Toggle Bold Print
\x05	Enquiry [ENQ]	User Print Command
\x0b	Vertical Tab [VT]	Odd/Even Page Offset
\x13	Device Control 3 [DC3]	Toggle Underline
\x14	NL Line Feed, New Line [LF]	Toggle Superscript
\x1a	Substitute [SUB]	End of File Marker
\x8d	Not Valid 8-bit Code [<control>]	Line Terminator on Word Wrap

3. INITIAL TOOL VALIDATION

After the initial content analysis we undertook an initial survey to identify potential tools for the conversion process. We found eight potential application/codecs; however, some simple trials very quickly indicated that none of the converters was able to accurately convert all the files in the set, to any format. The initial testing included two reference files being converted, and of the eight tools tested, only three managed to return even a valid file that could be measured against the original.

Given that we abandoned the testing of any commercially or otherwise available conversion product due to their inability to perform on our test files, the purpose of this paper is not to discuss the various functions and failings of each of the tools, suffice it to say, we could not find a single product that was able to offer us the ability to accurately convert our files.

Exploring these converters led to a complex part of the problem. How can we compare the accurate conversion of file A to a new format? What metrics can we use to convince ourselves of the efficacy of any conversion process?

At this early stage, it was enough to use simple word/character counts as the pass/fail metric. Each WordStar file was stripped of any non-printable UTF-8 characters, and each word/character was tallied. Once converted, and irrespective of the output format of the converter, the resulting text was also stripped of markup and each word/character tallied. These measures were collected and tables such as below were generated (Table 2 and Table 3).

Table 2 - File content metrics for file Ref1.WS

Ref_Name	Ref1.WS	App 1	Exact?
Pages	6	5	FALSE
Words	3639	3639	TRUE
Chars (no spaces)	15906	15906	TRUE
Chars(spaces)	19774	19515	FALSE
Paras	319	32	FALSE
Lines	328	222	FALSE

Table 3 - File content metrics for file Ref2.WS

Ref_Name	Ref2.WS	App 1	Exact?
Pages	34	32	FALSE
Words	24880	24880	TRUE
Chars (no spaces)	95005	95005	TRUE
Chars(spaces)	122530	119666	FALSE
Paras	1903	361	FALSE
Lines	1954	1434	FALSE

The second phase of testing was to try all the available WordStar files against the three known working tools to ensure that all of the WordStar files could be converted by the tools.

This phase highlighted the inconsistent and unsatisfactory performance of the tools, and a decision was made to write a new converter from scratch.

It is worth noting that the inconsistent performance encountered was largely found to be either a complete failure of the tool to return something usable as a converted file, or an inability to deal with all the variants of WordStar files found in the set.

4. CURATORIAL ENGAGEMENT

Having completed the initial trawl through the set and having a broad understanding of what was possible / viable, the next step was to ensure that curatorial concerns were fully understood.

An assessment template was created that formed the basis for a series of meetings with preservation and curatorial colleagues to ensure that the files under inspection were properly understood and thus properly migrated to a new format.

This assessment allowed the documents to be conceptually broken down into the various formatting and aesthetic features that comprise the original intellectual object and ensure that effort is expended in the right areas.

4.1 Working through the Original Content Review

Any preservation work undertaken by the NLNZ digital preservation team must be ratified by the "content owner". The content owner is the person within the Library who has overall responsibility for the collection items being preserved. The content in question are all transcripts from oral history recordings, so in this case, the owner is the Curator of Oral History and Sound. In addition, we enlisted the help of the digital archivist and assistant digital archivist to act as mediators between the Curator's content expertise and our digital expertise. They also gave valuable insights based on their own experiences.

Once there was agreement on the collection to be used, the next step was to undertake a review of the original content that could be used to measure the success (or not) of any proposed transformations.

Preservation planning must demonstrate that all aspects of the content has been considered and report on those aspects across the transformation. The discussions therefore focused not just on what must remain the same (significant properties), but just as

[6] This information was originally reverse engineered through inspection, and confirmed at a later date when the supporting WordStar (v3) manual [5] was discovered.

importantly, what could change: the measurement of a successful transformation must equally show what has and has not changed.[7]

The form we used looked at the mark-up and the formatting of the WordStar content. We could find no previous work on WordStar conversions that would aid this work, but did use various sources such as proof-reading notes for the types of mark-up and text-based features that we should look out for. The process of going through this form was deliberately elongated. This is the first "in-anger" preservation action that the Library has undertaken and we wanted to ensure that we were covering every eventuality and more importantly that the curator was entirely comfortable with what was being done to content. This would give us all confidence that the new representations of the content could stand as a true and accurate replacement for the original.

Our technique was to go through every possible aspect of the WordStar files and discuss their importance. The mutability of each aspect was the key currency for this conversation, for example; could the font-type change? What about margins in the documents? As we went through these twenty six identified aspects two primary issues arose:

1. What *facts* do we know about the original content?
2. Are we replicating the content seen on screen or as it was printed?

The first question speaks to the notion that we were not running equipment completely contemporaneous with the content. The version of WordStar used as our reference (WordStar for MSDOS v5) post-dates when we believe the transcripts were written. All decisions therefore are founded on the fact that we are not viewing the material in a completely original environment.[8]

The second question was only one that the curator could answer. What exactly were we trying to preserve? Is it the look and feel on the screen, including its markup? Or were we in fact preserving the output of the word processed document? We know that these files were created in order to be printed and given to those listening to the original recordings. It was decided that we were preserving how the content would have looked when it was printed.

The medium of presentation was not of any real concern. The Library made the decision a long-time ago that we are not a computer museum. We do not preserve the original hardware to present content nor do we wish to preserve content so that interaction with it is exactly as it was two decades ago. We always plan to represent the content with best efforts to retain original features, but always with an eye to allowing new and future use of it. That is to say, in this case, we were unconcerned with preserving the blue screen with the markup presented to the writer. We were more concerned with presenting the text in a fashion that the creator would recognise as their work. The content is king, not the medium.

4.2 What is a viable metric?

With any automated process, the question of measuring quality / effectiveness of any migration action arose as a concern. The WordStar documents were no different in this regard. Essentially, there is a need to find a "middle ground metric" that would allow the original file to be compared with the new files, and some automated decision making / logging to ensure that migration actions are accurate.

To simplify this process, the WordStar content was conceptually sliced into two concerns; aesthetic construction and intellectual content.

Each concern is considered in isolation of the other, and each has its own pass/fail measures that once satisfied will result in the final outcome fulfilling all concerns.

4.3 Concern 1: Aesthetic construction

This measure is essentially the visual appearance of the intellectual object. It cares not what the content "says"; it cares about capturing the "look and feel" of the original item.

Ostensibly this appears to concern itself with font, and text size. However, there is a deeper layer of considerations that address page layout and any stylistic application used in the document to convey intellectual concepts. In this case, the new paragraph indentation is regarded as the aesthetic evidence of the intellectual concept of "a paragraph" and the underline font used to denote speaker and time from the spoken words.

The discussion can be summarised as follows:

- of the files in the set, none have an explicit font type specified in the file object;
- of the files in the set, none have an explicit text / font size specified in the file object;
- the reference version of WordStar has a default font applied to any new document;
- this is assumed[9] to be a common feature of any version of WordStar;
- the reference version of WordStar has a default text / font size applied to any new document;
- this is assumed to be a common feature of any version of WordStar;
- unless explicitly stated, the font type, and text size used is assumed to be the default set by WordStar;
- unless explicitly stated / advised, the default font is taken to be "Courier";
- unless explicitly stated / advised, the default text size is taken to be 10 points.

[7] We have previously noted that we do not believe the current definition of significant properties is sufficient. The definition states that they are properties "determined to be important to maintain through preservation actions" [1]. Our opinion is that "all technical properties … [are] important, irrespective of whether or not they should remain across an action. Some properties we may actually want to deliberately take action on to remove from the file. These properties are significant and must be tracked across actions" [2].

[8] We work from the generalised position that you cannot view content in an environment that exactly matches the original environment (other than perhaps the original itself). There are too many unknown and known variables that can never be replicated to support the perfect rebuild of a system in its original environment.

[9] The word "assume" will trigger alarm bells for preservation specialists. We **must** make assumptions when we have exhausted other possibilities or else we would not be able to complete this work. We are comfortable with making assumptions as long as they are noted, consistent and based on a degree of contextual knowledge that must be used in the absence of any evidence to the contrary.

The full discussion is conveyed in the original content review outputs.

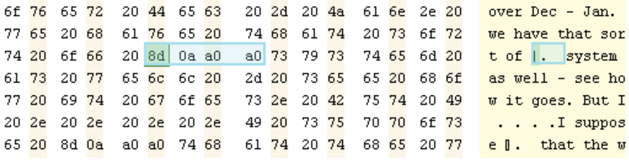
4.4 Concern 2: Intellectual content

In this measure, we are interested to ensure that all intellectual content accurately travels from file A to the migrated file B. There should be no translation of information, concepts or semantic-laden parts of the original file – only direct, absolute migration of content.

In principle, this seems a simple premise. However, there was need for significant discussion to ensure that we collectively understood the significance of various parts of the document.

One of the most interesting and far reaching discussions was on the purpose of counting "lines".

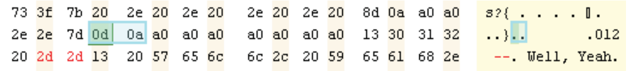
In WordStar, in the text editor, line endings and carriage returns are automatically inserted where required by the software. Lines appear on screen to be essentially fluid (a change at the top of a paragraph propagates line changes where needed along every line in the paragraph, as fitting within page margins). However, inspection of the file shows that these line endings are hard written into each line (using the hexadecimal marker \x8d\x0a\xa0\x0a):



```
6f 76 65 72  20 44 65 63   20 2d 20 4a  61 6e 2e 20   over Dec - Jan.
77 65 20 68  61 76 65 20   74 68 61 74  20 73 6f 72   we have that sor
74 20 6f 66  20 8d 0a a0   a0 73 79 73  74 65 6d 20   t of ▯.  system
61 73 20 77  65 6c 6c 20   2d 20 73 65  65 20 68 6f   as well - see ho
77 20 69 74  20 67 6f 65   73 2e 20 42  75 74 20 49   w it goes. But I
20 2e 20 2e  20 2e 20 2e   49 20 73 75  70 70 6f 73    . . . .I suppos
65 20 8d 0a  a0 a0 74 68   61 74 20 74  68 65 20 77   e ▯.  that the w
```

Figure 8 - Example word wrapped line ending

Hard typed carriage returns / line feeds (i.e. application of the "enter" key inside a text document) are indicated differently in the file stream to these "soft" line returns, (and are as expected in standard text documents \x0d\x0a):



```
 73 3f 7b 20  2e 20 2e 20   2e 20 2e 20  8d 0a a0 a0   s?{ . . . . ▯.
 2e 2e 7d 0d  0a a0 a0 a0   a0 a0 a0 a0  13 30 31 32   ..}..       .012
 20 2d 2d 13  20 57 65 6c   6c 2c 20 59  65 61 68 2e    --. Well, Yeah.
```

Figure 9 - Example explicit line ending[10]

From analysis of the corpus, it is apparent that some files have very differing page margins. This can be seen by making a histogram of the line lengths found in the files as a set (Figure 9).

The double peak is particularly interesting. If all documents had the same line margin, this would be observed as a single peak, as is found in most of the individual file analysis (see Figure 10).

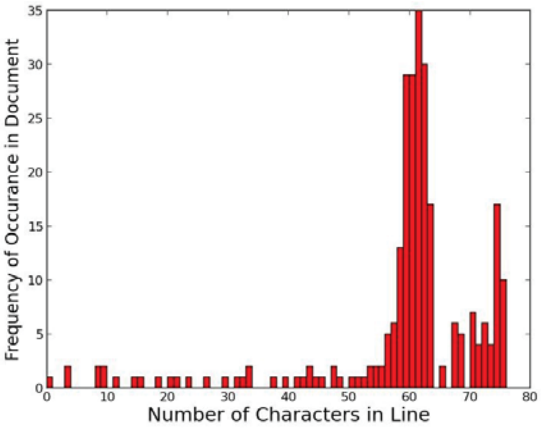


Figure 7 - Frequency of line length: All files in Corpus

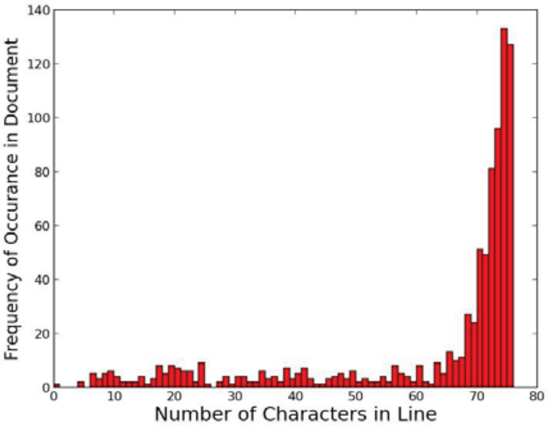


Figure 10 - Frequency of line length - File reference: #3

However, some files clearly show this double peak (Figure 11)

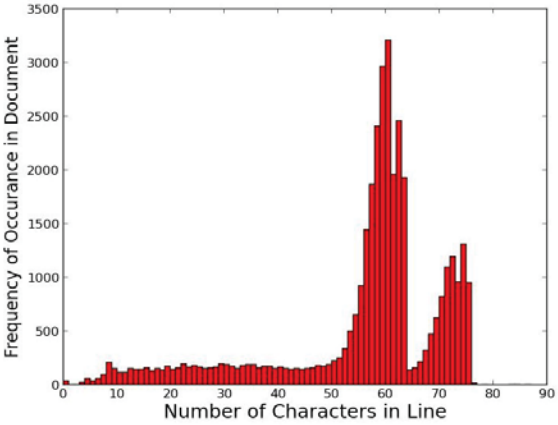


Figure 11 - Frequency of line length - File reference: #22

[10] Red text is a manual redaction of identifying names, places or initials that are found in the original document. This redaction method will be used through this paper.

In these files, it is notable that the line length varies in discrete "chunks" in the document, with no apparent explanation for the variation.

The ensuing conversation resolved that the concept of paragraph was primary, and there was no need to attempt to preserve the original line size, as this would likely impact negatively on the modern consumption of the intellectual object.

By deciding on this matter, the working group had essentially agreed that when the document was created, the original author had no explicit desire to reflect any meaning from the line length used. The length of a line was simply a by-product of the paragraph structure. In other cases, with other collections, this perhaps would not be a safe assumption and serves to reflect the importance of having curatorial involvement in the process.

This decision on the line-endings had an impact: the number of pages changes. We had to consider therefore if people had referenced these documents and how they did so. Did they (or would they in future) reference by page number? The decision was made that in this case, the movement of text across pages was allowable as accurate reference would be made through time-points noted in the text rather than page numbers. However, it was an impact that required some considerable attention.

Some other decisions made can be summarised. The use of underlined and bold fonts in the document was seen as of intellectual import and as such should be perfectly replicated in the migrated final. All standard text characters should be migrated with no change. Paragraph structure is essential to replicate accurately, original line length is not. The paragraph object and the "word" (an ordered group of printable characters) are considered the primary intellectual concerns to migrate and measure.

The second intellectual concept to consider is the conventions used by the author of the document to convey informational components. In this set, we are fortunate that all the documents come from the same source, and as such share a common set of conventions. These were noted as:-

- A new paragraph is indented on the page
- A speaker is denoted by their initials and these are underlined. E.g. JG
    - These are occasionally bolded. E.g. **JG**
- Time elapsed in the interview is recorded as an underlined number. E.g. 005
    - These are occasionally bolded. E.g. **005**
- On occasions these features are combined with at least one white space char as a separator, E.g. JG 005
    - These are occasionally bolded. E.g. **JG 005**
    - Order is not controlled, both JG 005 and 005 JG are found in the texts

This means that we have up to three intellectual concepts that are clearly identifiable in each paragraph; a speaker, a chronological marker, and the spoken words. These features were to be retained.

These pieces of information (the aesthetic and intellectual pieces) formed the backbone of proofs that were presented to the content-owner in the preservation plan.

5. WRITING THE CONVERTR

Having spent time with the content files and the content curator ensuring that the WordStar files were well understood (technologically, intellectually and aesthetically) the next step was to choose a target format, and construct the converter.

5.1 Picking a target format

Given that these files are known to be relatively simple text only files, but do contain some basic formatting, we could rule out some formats and rule in some others.

The master list of options included any valid variant of PDF, MS DOC (OLE2 based), MS DOCX, RTF, ODF text variant or HTML (v4 or v5).

We already have content in all these formats, creating more of any of these would be viable, however we had to choose one, and that decision was made against the following points:

PDF – High ranked candidate. Can be problematic if not properly constructed.

MS DOC – Not ideal, proprietary standard. We would need a specific encoding library to create valid doc files.

MS DOCX – As above, but slightly more preferred due to the availability of its specification.

RTF – Not ideal. Known to be problematic at times if not implemented well.[11] Not well suited in our current delivery environment.[12]

ODF – Not ideal. Not widely used / supported / found in collections.

HTML – High ranked candidate (if all formatting requirements are supported). Easy to use, easy to create, open standards. Relatively transparent.

HTML v4 was picked as target new format. The main justifications were:

- very common open standard;
- very well supported standard;
- found in significant volume in the collections;
- supports the formatting requirements;
- easy to wrangle into preferred shape;
- results in low complexity files.[13]

5.2 Writing and testing the converter

The target language for the converter was Python. This was a natural choice as it has native support for text manipulation.

This was one of the first Python projects ever completed by the code developer – and as such it should be noted that the code used is not always the most efficient / simple / pythonic implementation.[14]

When planning the build, the process was broken into some core tasks.

[11] NLNZ has previously undertaken remediation work on RTF files prior to ingest

[12] Ease of access is one criteria in our preservation planning process.

[13] The full discussion of the merits of the formats is contained in the preservation plan.

[14] Examples of the code are given in Appendix 1.

1. Take WordStar file
2. Slice file into paragraphs
3. Per paragraph
    a. Strip formatting, make text only version for comparison
    b. Convert WordStar markup into HTML markup
    c. Recombine into a document
4. Apply HTML structure
5. Save new file
6. Open new file
    a. Strip formatting, make text only version for comparison
7. Compare reference versions with each other
8. Write log
9. End

During the conversion a number of challenges arose. One was based on the fact that WordStar markup tags do not contain "start" or "stop" information. They are a simple binary switch, or a "toggle". For example, bold is either turned on, or turned off.

Conversely, HTML tags do contain "start" or "stop" information. <b> de-marks the start of bold text and </b> indicate the end of bold text.

This poses two interesting challenges.

- What happens when an author fails to close the bold tag?
- As we have introduced the concept of a paragraph as a structural object, and must comply with HTML rules for nesting elements, what happens if a tag pair:

  ```
  (<b>Some arbitrary text</b>)
  ```

  overruns a paragraph boundary?

In the migration code above, the parser looks for the bold tag \x13 and if detected, it flips the bold_marker flag. If it was False, it becomes True, (and visa versa), and inserts the corresponding tag into the text. At the end of the paragraph the decision was made to force any open tags to close. This prevents the formatting from "leaking" into the rest of the document when it's not closed properly due to an author omission.

A secondary issue emerged once all the formatting tags were implemented. HTML is very deliberate about tag order. Tags are expected to open and close in order.

For example,

```
<b><u>Some arbitrary text</b></u>
```

would not be valid HTML,

```
<b><u>Some arbitrary text</u></b>
```

would be. Note the positioning of the closing tags. This meant that tag nesting and detection of invalid tag sets was required to ensure that valid HTML was generated.

5.2.1 Conversion Principle Development

As the conversion code was tested and iterated, a conversion principle was refined. Namely that the conversion code should only emulate the performance and behavior of the original software, and not address any formatting errors viewed to be editorial. This became a useful principle to lean on as the conversion code became more complete.

This principle was tested at length when considering how to handle multiple spaces in a document. Analysis of the document corpus showed the number of times the author used white space in a way that would be suppressed by HTML unless mitigated:

- Double space between two printable chars: 372
- Triple space between two printable chars: 113
- Between 2 and 50 spaces between two printable chars: 870
- A single space between a printable char and a full-stop: 3,992

As HTML is not a whitespace preserving format, whitespace ranges of longer than one character would need to be processed in such a way that the browser was forced to render each character, and not to conflate spans of whitespace into a single character.

This was achieved by converting whitespace spans longer than one character to a mixture of breaking and non-breaking white space characters. The non-breaking white space character ("\xc2\xa0") is always rendered by the browser or HTML parser, and so was used to ensure that white space characters were reproduced exactly as per the original.

This was particularly important where whitespace was used between formatting tags.

For example, the WordStar section

```
 ^U This is some arbitrary text ^U
```

would natively convert to HTML as:-

```
<p> <u> This is some arbitrary text </u><p>
```

which in turn would render as:-

"This is some arbitrary text"

The actual expected WordStar formatted text should render as

" This is some arbitrary text "

Note the leading and trailing whitespace.

It was therefore important that this was handled correctly to ensure that the formatting as was found in the original files was accurately moved into the HTML files.

5.2.2 Addressing the aesthetic concerns

As previously discussed, there was much analysis of the aesthetic construction of the files. This concerned the font choice, font size, line width, and the various page margins.

In the WordStar files, the font selection, and page margins are defined primarily in the default template. It is possible for an author to manually change these values. However, this would have left something of a footprint in the files and was not detected.

A decision was made to use an internal CSS declaration in the HTML documents to declare the font, font size and margins. The font was set to "courier" and the margins adjusted to ensure that the page layout follows the norms found in the original files.

This allowed the "speech" line starts to be indented as per the originals, and the wrapped lines to be pulled away from the edge of the HTML frame replicating the margin found in the original.

The CSS declaration used was:-

```
<STYLE TYPE="text/css">
```

```
        <!--
        BODY { margin: 1px 1px; text-indent:-2em;
font-family:courier; }
        p { text-indent:-2em; padding-left: 2em;
margin:4px 0px; }
        -->
</STYLE>
```

5.2.3 *Dealing with exceptions*

The main exception was a single line of formatting found in one document.

The line of interest was:

`\xc2\xa0\xc2\xa0\xc2\xa0\xc2\xa0\xc2\xa0\xc2\xa0\xc2\xa0\xc2\xa0\x14\x05\x14But I mean, when you say there are lots of contradictions, there \xc2\x8d`

Breaking this line down gives:

`\xc2\xa0\xc2\xa0\xc2\xa0\xc2\xa0\xc2\xa0\xc2\xa0\xc2\xa0\xc2\xa0` (The normal indentation)

`\x14\x05\x14` (The problem area)

`But I mean, when you say there are lots of contradictions, there` (The "text")

`\xc2\x8d` (The normal hard coded line wrap)

It is worth noting the context of the line. It comes from a longer piece of speech by one of the speakers and so is found "mid flow" and the problem bytes are unique to this line, in this file, in the corpus.

The byte `\x14` is used by WordStar to denote superscript text, (toggling on and off as per other formatting markers). The byte `\x05` is used by WordStar to support user generated codes to be sent to the printer directly.

This has some interesting connotations. Because the original install is not available for inspection and there are no supporting notes, it is impossible to know what code might have been used here. This would have been a printer specific instruction and set up in the installed instance of WordStar by the user. The version of WordStar these files were created on supports up to four of these user codes and so even if the printer and the possible codes that could have been used were known, the specific code bound to the byte is long forgotten.

The combination of bytes essentially says:

<toggle superscript on>

<send unknown user code to printer>

<toggle superscript off>

Given that any user code is unknown and its impact on the document is impossible to guess, it was decided to remove both these bytes ("toggle superscript", and "send user code to printer") from the document. The justification being that it was either an error by the author, or that its impact cannot be sensibly guessed. There is no known visible text inside the superscript tags and when printed on the reference build of WordStar, it had no affect on the printed text.

5.3 Building a text comparison tool

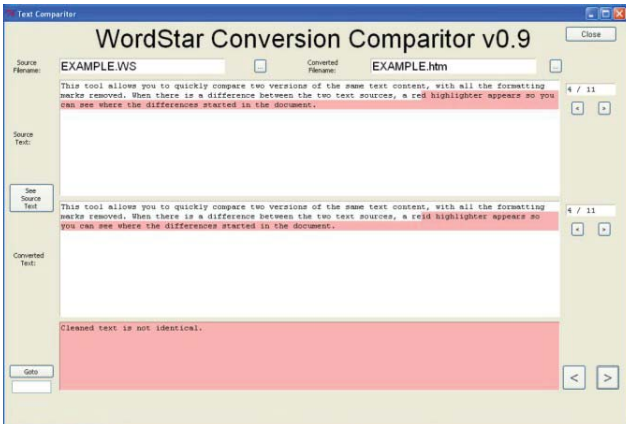
Having agreed on the aesthetic treatment of files, there was an outstanding question of how the intellectual accuracy could be demonstrated to the curator. The conversion code never actually "touches" normally encoded text characters; it simply moves them into the new HTML file. In the conversion process, a word by word check is made to ensure this is true and a log generated to record this fact.

It was undesirable to require the curator colleagues to read the conversion script and produced file to assure themselves that every word was there. A decision was made therefore to build a simple text comparison tool to allow the curator to inspect the new file, comparing it with the original.

The comparison tool was built in python and was designed to allow a reader to step through a file, paragraph by paragraph. The tool displays the original paragraph the new HTML paragraph, and a summary of any differences found in the use of alpha numeric characters, punctuation and whitespace.

The reader was able to toggle between a "cleaned" version of the text (with all formatting removed) and the native paragraph as found in each file. It displayed filenames, and paragraph numbers to allow any discrepancies to be recorded and later investigated.

Some basic navigation tools were included (such as "jump to paragraph number n") and key bindings to allow any file pair, and their associated text parts to be swiftly assessed.



EXAMPLE.WS

EXAMPLE.htm

This tool allows you to quickly compare two versions of the same text content, with all the formatting marks removed. When there is a difference between the two text sources, a red highlighter appears so you can see where the differences started in the document.

4 / 11

This tool allows you to quickly compare two versions of the same text content, with all the formatting marks removed. When there is a difference between the two text sources, a reid highlighter appears so you can see where the differences started in the document.

4 / 11

Cleaned text is not identical.

Figure 12- Screen shot of the comparison tool

This proved to be an invaluable tool in allowing the curator to demonstrate to their satisfaction that the files were accurately converted.

The tool allowed the curator to spend time with the content, at their own pace, assessing the original files in a meaningful way, and comparing the proposed conversions. They were able to see behind the relatively dense conversion code, and look at the raw information found in the source files being migrated.

6. LESSONS LEARNED

Because of the exploratory nature of the project, the end to end process took a long time to complete. Each step was very carefully considered by technical and curatorial staff alike, and it was deemed valuable to explore every question or concern in detail when it was encountered.

It would be unreasonable to attempt to calculate the amount of effort that went into completing this work, not least because one of the stated aims of the project was to give us the time and space to explore the concept of migration, to develop key skills in this area, build tools, wrangle files and otherwise build a strong foundation to help support our broad program of work.

The first lesson was one of comfort. Whilst the number of files in the corpus being converted was low, the methodical and thorough nature of the assessments, as presented in the preservation plan, resulted in a strong comfort that the conversion was accurate, thus satisfying the curator. Further to this point, the way our preservation system is designed means that the original WordStar files are never actually replaced by the HTML files in the system. They are superseded in the versioning model used to describe the intellectual entity. This means that if better tools were developed for this migration, it would be a trivial exercise to return to the WordStar originals and make new converted versions directly from the original content.

Managing expectations was critical in creating the environment for all actors to be comfortable with the results. We are not trying to recreate absolutely the original, but rather create a version of the original content that can stand in the original's stead and allow use and reuse of that content.

The second lesson was that there should be no assumptions about the context and knowledge that colleagues bring to the process. During the work a surprising paradigm shift was made by project workers. In the early exploration of conversion tools, any suggestion that converted artifacts might be further processed beyond what the tool had already done to produce more accurate results was met with stern a stern "no". That "no" described an unfamiliarity with methods of conversion and the separation of content from medium. In the early stages of the project, before we created our own tools, a suggestion was mooted that a line could be added to the HMTL files (once the conversion utility had finished with them) that would lock the font as courier, rather than allowing the browser to choose the font. This was met with some resistance. The main argument used was that the HTML would have to be changed, resulting in the HTML created by the commercial conversion tool being somehow "disrupted".

The counter argument was that this should not be an issue. The commercial conversion tool was effectively a black box, and there was no deep knowledge of what was happening inside the tool to create the converted files. To that end, it should not be problematic to make some known changes to the files (the adding of the defined font) when the rest of the processing used by the tool was unknown. Of course, this argument became irrelevant once the decision was made to build a conversion tool from scratch, however at the time it was an interesting point to explore.

The third lesson was figuring out how to succinctly demonstrate technical processes to non technical colleagues. Some of the decision making was very technical but required the support and validation of non technical colleagues. We used the original content review as a method of framing these discussions and deliberately took the time to ensure that a full understanding was achieved. In the future, and with other curators, it may not be necessary to take them on the entire technical journey. In this case though, we felt the time and effort taken to build a good relationship with the curator was important to ensure acceptance of what we were proposing. It also had the added advantage of tightening our own understanding of the processes and attitudes towards them.

By working to the agreed principles and making simple tools that allowed technical processes to be easily demonstrated, it was possible to put the right level of detail in the hands of decision makers to enable them to understand what was happening at all times during the project. Ultimately, education on both sides of the technical divide took place across the entire process.

7. CONCLUSIONS AND NEXT STEPS

At the culmination of this project, we are satisfied that we achieved the two aims we set out with.

The first aim was to explore the migration process for the Library, and get a sense of the complexity we face when attempting to move content from one format to another. This process needed to be robust, thorough, and transparent.

We noted that it took far longer than we would normally expect, and we are happy that the time taken we needed to ensure that all involved parties had the time and space to understand what we needed to do, and could contribute to the process in a meaningful way. We have identified some areas that can be significantly expedited and remain confident that the next iteration of this process would take less than half the effort we expended on this project.

Secondly we are confident that we have successfully moved the WordStar content in to the HTML format in an accurate and transparent way.

7.1 Next Steps

As a direct outcome of this project we will start to process our WordStar2000 content in a similar way. This content is related to the corpus addressed in this project, but different in its technical composition.

The learnings and tools from this project will be leveraged to deliver this next migration.

Given the very bespoke nature of the resulting conversion code, we do not plan to release the code as a supported application, or as "abandonware". The risk that it is used on content without the appropriate amount of technical / curatorial assessment is a liability we do not wish to hold. However, the code can be requested from the Library / paper authors, who would be happy to oblige.

8. ACKNOWLEDGMENTS

Our thanks to Linda Evans (Curator Oral History and Sound) Leigh Rosin (Digital Archivist) and Jessica Moran (Assistant Digital Archivist) for working with us through the process of migrating these files.

9. References

[1] PREMIS Editorial Committee, PREMIS Data Dictionary for Preservation Metadata, version 2.1, (January 2011), pg. 39.

[2] Jailani, H., McKinney, P. 2012. 'Compliance Conundrums: Implementing PREMIS at two National Libraries', *Proceedings of IS&T Archiving 2012*, Copenhagen.

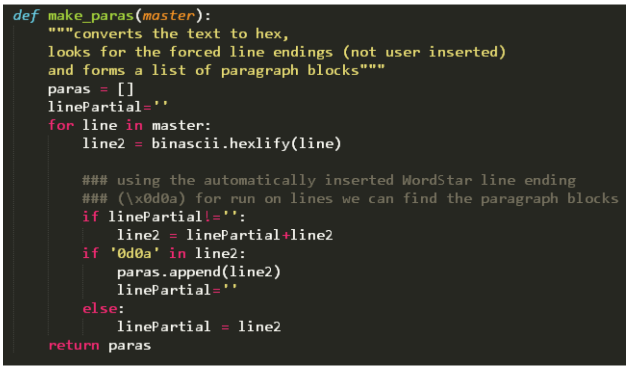
[3] The National Archives, *PRONOM Technical registry*, 31st March 2014. http://www.nationalarchives.gov.uk/PRONOM/.

[4] DeVorsey, K., McKinney P., 2010. 'Digital Preservation in Capable Hands: Taking Control of Risk Assessment at the National Library of New Zealand', *Information Standards Quarterly,* Spring 2010, 22:2, pp 41-44. http://www.niso.org/apps/group_public/download.php/4242/IP_DeVorsey_McKinney__Risk_Assessment_isqv22no2.pdf.

[5] MicroPro International Corporation, 1981. *WordStar Reference Manual.*

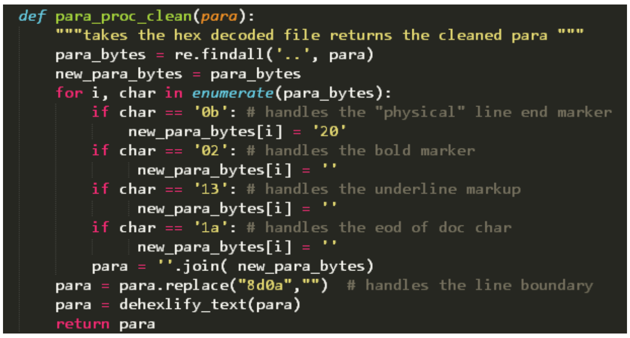
[6] The Internet Engineering Taskforce, 2003. *RFC 3629. UTF-8, a transformation format of ISO 10646,* http://www.ietf.org/rfc/rfc3629.txt.

10. Appendix 1: Examples of code



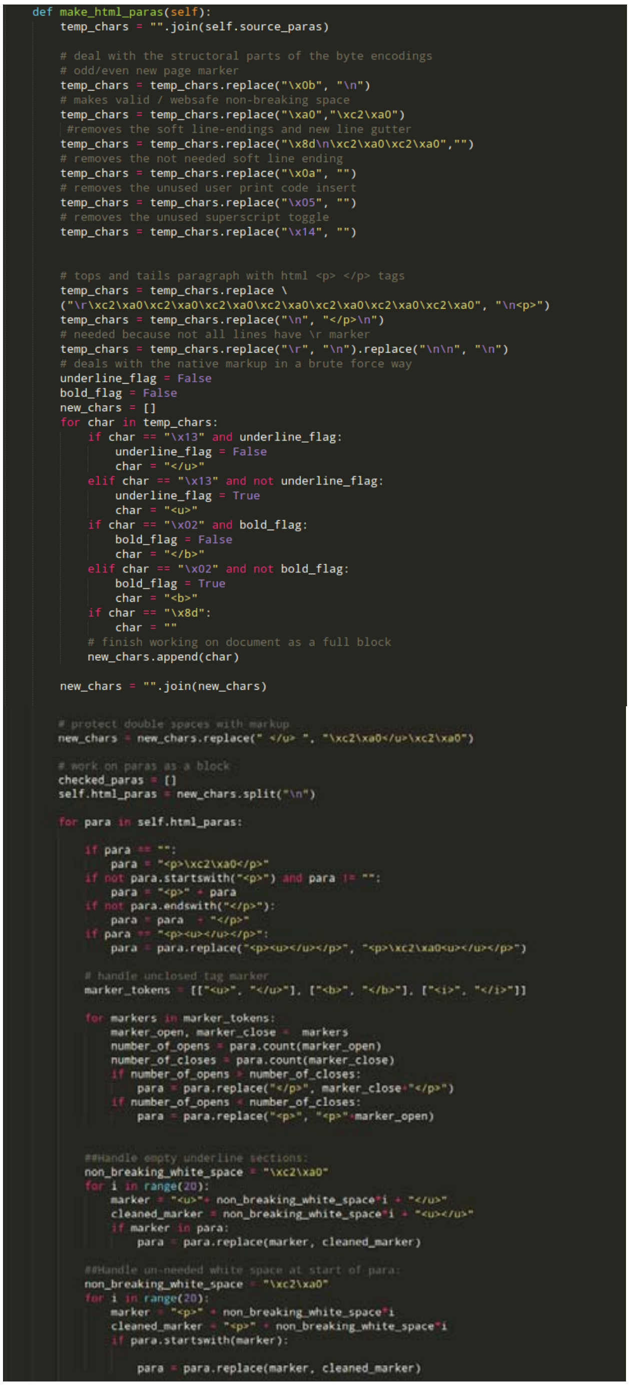
```python
def make_paras(master):
    """converts the text to hex,
    looks for the forced line endings (not user inserted)
    and forms a list of paragraph blocks"""
    paras = []
    linePartial=''
    for line in master:
        line2 = binascii.hexlify(line)

        ### using the automatically inserted WordStar line ending
        ### (\x0d0a) for run on lines we can find the paragraph blocks
        if linePartial!='':
            line2 = linePartial+line2
        if '0d0a' in line2:
            paras.append(line2)
            linePartial=''
        else:
            linePartial = line2
    return paras
```

Figure 13 – Split text into paragraphs

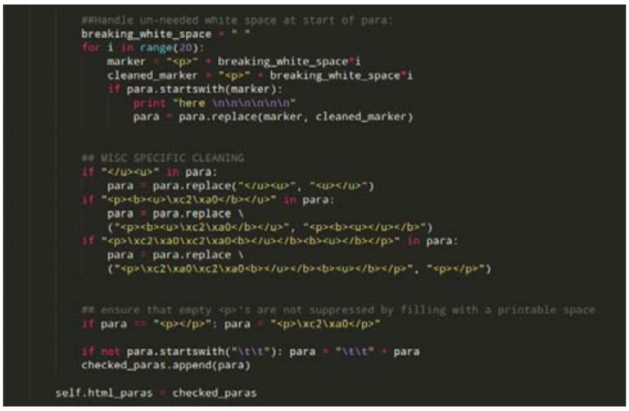


```python
def para_proc_clean(para):
    """takes the hex decoded file returns the cleaned para """
    para_bytes = re.findall('..', para)
    new_para_bytes = para_bytes
    for i, char in enumerate(para_bytes):
        if char == '0b': # handles the "physical" line end marker
            new_para_bytes[i] = '20'
        if char == '02': # handles the bold marker
            new_para_bytes[i] = ''
        if char == '13': # handles the underline markup
            new_para_bytes[i] = ''
        if char == '1a': # handles the eod of doc char
            new_para_bytes[i] = ''
        para = ''.join( new_para_bytes)
    para = para.replace("8d0a","")  # handles the line boundary
    para = dehexlify_text(para)
    return para
```

Figure 14 - Making plain text (no formatting)



```python
def make_html_paras(self):
    temp_chars = "".join(self.source_paras)

    # deal with the structoral parts of the byte encodings
    # odd/even new page marker
    temp_chars = temp_chars.replace("\x0b", "\n")
    # makes valid / websafe non-breaking space
    temp_chars = temp_chars.replace("\xa0","\xc2\xa0")
     #removes the soft line-endings and new line gutter
    temp_chars = temp_chars.replace("\x8d\n\xc2\xa0\xc2\xa0","")
    # removes the not needed soft line ending
    temp_chars = temp_chars.replace("\x0a", "")
    # removes the unused user print code insert
    temp_chars = temp_chars.replace("\x05", "")
    # removes the unused superscript toggle
    temp_chars = temp_chars.replace("\x14", "")


    # tops and tails paragraph with html <p> </p> tags
    temp_chars = temp_chars.replace \
    ("\r\xc2\xa0\xc2\xa0\xc2\xa0\xc2\xa0\xc2\xa0\xc2\xa0\xc2\xa0", "\n<p>")
    temp_chars = temp_chars.replace("\n", "</p>\n")
    # needed because not all lines have \r marker
    temp_chars = temp_chars.replace("\r", "\n").replace("\n\n", "\n")
    # deals with the native markup in a brute force way
    underline_flag = False
    bold_flag = False
    new_chars = []
    for char in temp_chars:
        if char == "\x13" and underline_flag:
            underline_flag = False
            char = "</u>"
        elif char == "\x13" and not underline_flag:
            underline_flag = True
            char = "<u>"
        if char == "\x02" and bold_flag:
            bold_flag = False
            char = "</b>"
        elif char == "\x02" and not bold_flag:
            bold_flag = True
            char = "<b>"
        if char == "\x8d":
            char = ""
        # finish working on document as a full block
        new_chars.append(char)

    new_chars = "".join(new_chars)

    # protect double spaces with markup
    new_chars = new_chars.replace(" </u> ", "\xc2\xa0</u>\xc2\xa0")

    # work on paras as a block
    checked_paras = []
    self.html_paras = new_chars.split("\n")

    for para in self.html_paras:

        if para == "":
            para = "<p>\xc2\xa0</p>"
        if not para.startswith("<p>") and para != "":
            para = "<p>" + para
        if not para.endswith("</p>"):
            para = para  + "</p>"
        if para == "<p><u></u></p>":
            para = para.replace("<p><u></u></p>", "<p>\xc2\xa0<u></u></p>")

        # handle unclosed tag marker
        marker_tokens = [["<u>", "</u>"], ["<b>", "</b>"], ["<i>", "</i>"]]

        for markers in marker_tokens:
            marker_open, marker_close =  markers
            number_of_opens = para.count(marker_open)
            number_of_closes = para.count(marker_close)
            if number_of_opens > number_of_closes:
                para = para.replace("</p>", marker_close+"</p>")
            if number_of_opens < number_of_closes:
                para = para.replace("<p>", "<p>"+marker_open)


        ##Handle empty underline sections:
        non_breaking_white_space = "\xc2\xa0"
        for i in range(20):
            marker = "<u>"+ non_breaking_white_space*i + "</u>"
            cleaned_marker = non_breaking_white_space*i + "<u></u>"
            if marker in para:
                para = para.replace(marker, cleaned_marker)

        ##Handle un-needed white space at start of para:
        non_breaking_white_space = "\xc2\xa0"
        for i in range(20):
            marker = "<p>" + non_breaking_white_space*i
            cleaned_marker = "<p>" + non_breaking_white_space*i
            if para.startswith(marker):

                para = para.replace(marker, cleaned_marker)
```



```
    ##Handle un-needed white space at start of para:
    breaking_white_space = " "
    for i in range(20):
        marker = "<p>" + breaking_white_space*i
        cleaned_marker = "<p>" + breaking_white_space*i
        if para.startswith(marker):
            print "here \n\n\n\n\n\n"
            para = para.replace(marker, cleaned_marker)


    ## MISC SPECIFIC CLEANING
    if "</u><u>" in para:
        para = para.replace("</u><u>", "<u></u>")
    if "<p><b><u>\xc2\xa0</b></u>" in para:
        para = para.replace \
        ("<p><b><u>\xc2\xa0</b></u>", "<p><b><u></u></b>")
    if "<p>\xc2\xa0\xc2\xa0<b></u></b><b><u></b></p>" in para:
        para = para.replace \
        ("<p>\xc2\xa0\xc2\xa0<b></u></b><b><u></b></p>", "<p></p>")


    ## ensure that empty <p>'s are not suppressed by filling with a printable space
    if para == "<p></p>": para = "<p>\xc2\xa0</p>"

    if not para.startswith("\t\t"): para = "\t\t" + para
    checked_paras.append(para)

self.html_paras = checked_paras
```

Figure 15 - Making HTML4 text

A Model for Format Endangerment Analysis using Fuzzy Logic

Roman Graf
AIT - Austrian Institute of Technology GmbH
Donau-City-Strasse 1
Vienna, Austria
roman.graf@ait.ac.at

Sergiu Gordea
AIT - Austrian Institute of Technology GmbH
Donau-City-Strasse 1
Vienna, Austria
sergiu.gordea@ait.ac.at

Heather Ryan
University of Denver
Library & Information Science Program
1999 E. Evans Avenue
Denver, CO 80208
heather.m.ryan@du.edu

ABSTRACT

This paper presents an approach for merging information automatically aggregated from open repositories and expert knowledge related to digital preservation. The main contribution of this work is the employment of fuzzy models to support digital preservation experts with semi-automatic estimation of "endangerment level" for file formats. Our goal is to make use of a solid knowledge base automatically aggregated from linked open data repositories to detect conflicts and inaccuracies in this data in order to improve the quality of a risk analysis process. The proposed method is meant to facilitate decision making with regard to preservation of digital content in libraries and archives using domain expert knowledge. To allow reasoning, even in the case of inconsistent data, we employ fuzzy logic techniques for transforming information about formats with user friendly metrics. The goal is to bring conflicting and incorrect information to the surface for correction and improvement by community. The analysis of a survey regarding the risk factors for file formats was used as an input for the fuzzy model and is presented in the evaluation section.

Categories and Subject Descriptors

H.3.7 [**Digital Libraries**]: System issues; H.3.5 [**Online Information Services**]: Web-based services

General Terms

infrastructure

Keywords

digital preservation, risk analysis, linked open data, preservation planning, ontology matching, information integration

1. INTRODUCTION

In recent years, libraries, archives and museums have been carrying out large-scale digitization projects and have been including an increasing amount of born digital content in their collections. As a result, new digital collections that comprise millions of objects were created; and the goal is to make them available on long term basis. Consequently, digital libraries are facing a paradigm shift regarding preservation, maintenance and quality assurance of these collections. Therefore, automated solutions for data management and digital preservation are imperatively necessary.

One of the core preservation activities deals with the evaluation of appropriate formats used for encoding digital content. The preservation risks for a particular file format are difficult to estimate [Graf and Gordea 2013]. The definition of risk factors and associated metrics is still an open research topic in the digital preservation community[1]. Involvement of digital preservation experts is required for collecting complete information and evaluating preservation risks[Ayris et al. 2008]. Currently, each institution defines its own risk factors for long term preservation depending on particular project, preservation goals, workflows and assets. The richness and the quality of individual knowledge bases play an important role in making decisions on preservation planning, but often these resources do not contain all of the necessary semantic information for performing a faithful (automatic) evaluation of file formats.

Many file formats are properly documented, are open-source and well supported by software vendors. Other formats may be outdated or no longer functional with modern software or hardware. There are also custom/proprietary formats, which might be obsolete and not renderable with commodity hardware. To address these problems, we employ the File Format Metadata Aggregator (FFMA) [Graf and Gordea 2012]) system and the information integration approach depicted in Figure 1. FFMA is a part of knowledge base recommender DiPRec [Gordea et al. 2011], which reuses the experience of building preservation planning tools and offers assessment for long-term preservation of digital content. This tool performs an analysis of file formats based on the concept of risk scores.

The main contribution of the current work is the development of an Expert System based on fuzzy rules for performing the analysis of digital collections. Fuzzy rules are employed for handling the level of uncertainty associated with the information aggregated from Linked Open Data (LOD). Decision support based on the elaborated rule engine provided by FFMA and fuzzy rules is meant to support institutions like libraries and archives with assessment for

[1] `http://www.openplanetsfoundation.org/blogs/2013-09-30-assessing-file-format-risks-searching-bigfoot`

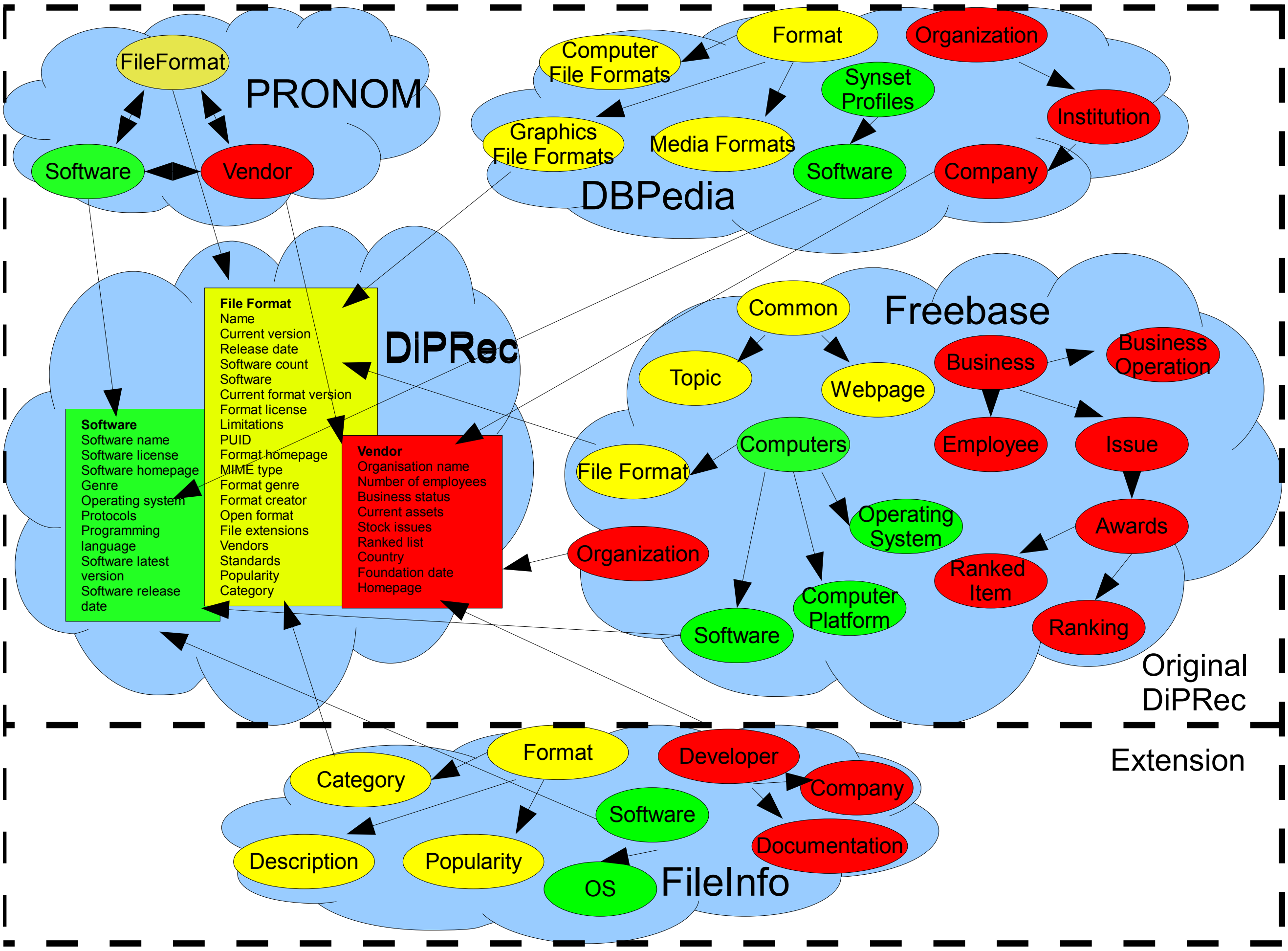


Figure 1: PRONOM, DBPedia, Freebase and Fileinfo digital preservation domain related ontology sections mapped to the DiPRec file format ontology.

analyzing their digital assets. The basis for risk metrics calculation was provided by study organised by Heather Ryan while she was at the University of North Carolina at Chapel Hill[Ryan 2013] which takes in account twenty eight risk factors. Evaluation metrics were defined for each of these factors based on the knowledge of digital preservation community. We aim at defining a fuzzy model and metrics intended to provide decision making support based on expert community knowledge. The paper is structured as follows: Section 2 gives an overview of related work and concepts. Section 3 explains the risk analysis process, knowledge aggregation process from LOD repositories as well as ontology mapping, fuzzy modelling and algorithmic details of endangerment analysis. Section 4 presents the experimental setup, file formats study, applied methods for fuzzy analysis and results. Section 5 concludes the paper and gives an outlook about planned future work.

2. RELATED WORK

The main issue addressed in this work is the controversial understanding of format obsolescence. Andrew Jackson provides an overview of this topic in [Jackson 2012] where he evaluated competing hypotheses regarding the software obsolescence issue. He employed format identification tools for selecting appropriate preservation strategies. One of these hypothesis is presented by Rothenberg [Rothenberg 2012] and emphasizes that all formats should be considered brittle and transient, and that frequent preservation actions will be required in order to keep data publicly accessible. In contrast to that hypothesis Rosenthal [Rosenthal 2010] claims that no one supporter of format migration strategy was able to identify even one format that has gone obsolete in the last two decades. Rosenthal argues that the network effects of data sharing inhibit obsolescence.

Accurate format identification and rendering is a challenging task due to malformed MIME types, rendering expenses, dependence on some content not embedded in the file, missing colour tables, changed fonts, etc. In [Jackson 2012], the author examines how the network effects could stabilise formats against obsolescence in order to understand the warnings, choices and costs involved. This evaluation should help to meet a preservation strategy: either to perform frequent preservation actions to keep data accessible or to concentrate on storing the content and using available rendering software. The result of evaluation demonstrates that most formats last much longer than five years, that network effects stabilise formats, and that new formats appear at a modest, manageable rate. However, he also found a number of formats and versions that are fading from use and that every corpus contains its own biases.

The digital preservation tools like PANIC [Hunter and Choudhury 2006], AONS II [Pearson and Webb 2008], SPOT [Vermaaten et al. 2012], P2 registry [David Tarrant 2011], aimed at identifying file formats used for encoding digital collections and informing repository managers of events that might impact the access to the stored content. They also define mechanisms for alerting when file formats become obsolete. These tools demonstrate significant differences to our approach. They do not apply metrics for risk calculation, and take in account significantly fewer properties. Often these properties are estimated and not measurable, do not exploit the knowledge available to the public, or are limited to particular open sources. Also, there is no common understanding in the community about the meaning of the term "obsolete" as mentioned above. In the proposed approach we do not intend to mark down obsoleted formats, since there are different hypotheses and no common accepted definition for format obsolescence. We estimate obsolescence in relation to the additional effort required to render a file beyond the capability of a regular PC setup in a particular institution. This is consistent with the "institutional obsolescence" concept saying that a particular format that would no longer render on a PC in an institution's reading room should be considered obsolete.

An application of Natural Language Processing (NLP) instead of numerical data for computing and reasoning using fuzzy logic is described in [Lee 1990]. A survey of the fuzzy logic controller (FLC) presented in [Zadeh 1996] evaluates a linguistic control methodologies, the derivation of the fuzzy control rules and an analysis of fuzzy reasoning mechanisms. The qualitative safety modelling in [Sii et al. 2001] is performed employing fuzzy IF - THEN rules. Compared to existing digital preservation recommenders the proposed approach is more effective due to the use of more complex fuzzy rules. Existing tools are not well suited for dealing with aggregated LOD data having a level of uncertainty due to conflicts and inaccuracies between different sources. Inaccuracies in this sense are slightly different measurements, which do not impact the overall evaluation of the risk factor. E.g. software count for PDF format provided by Freebase is 12 whereas Fileinfo describes 25 tools. We define conflicts as significant contradictions implying different conclusions on risk factor evaluation. E.g. PRONOM classification for PDF format is "page description" that contradicts the Freebase genre for this format, "graphics file format". A fuzzy-logic-based approach is more appropriate for the correctness analysis. The provided Expert System deals directly with the linguistic terms commonly used in the digital preservation community for quality assessment. Our research focuses on the development and representation of user friendly and easily understandable linguistic variables to confidence levels. These variables are then quantified using fuzzy logic. Inspired by [Pearson and Webb 2008] we realized the need to develop a central web service that shares the results of open data aggregation and correctness assessments with the community of interest. We aim at defining endangerment metrics based on the experience of community members who share their individual expertise on defining and identifying risk factors.

3. ENDANGERMENT ANALYSIS

Digital preservation is an area where we have to take into account fuzziness and a high amount of descriptions regarding the encoding formats. The description of file formats aggregated from open repositories is often far from being complete and accurate. Therefore, we support the aggregation of expert knowledge for enhancing such a repository with high confidence information. The proposed Expert System should identify conflicts and inaccuracies and provide assessment on the "institutional obsolescence" of file formats. We realized that the digital preservation community already uses multiple format registries and doesn't trust "expert systems" for making preservation related decisions. Instead, they recognize the need for support systems that aggregate and compare knowledge about the file formats (i.e. in form of metrics). This approach should help to uncover conflicting and untrusted information so that domain experts may correct it according to the policies established in their institution.

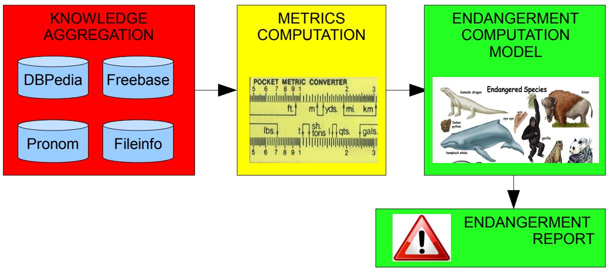


Figure 2: The workflow for the format endangerment analysis.

Figure 2 sketches the workflow used within the endangerment analysis process. The creation of endangerment analysis reports is a two-step process based on the definition of fuzzy factors (i.e. Endangerment Computation Model). The second step is the computation and interpretation of fuzzy metrics (i.e. Metrics Computation). The building of the knowledge base (i.e. Knowledge Aggregation) is a prerequisite for performing the endangerment computations[Graf and Gordea 2013]. This includes the acquisition of expert knowledge and the aggregation of file format data in a common domain model. The final report contains detailed information about the endangerment level, including quantifications of the evaluation factors, the computed metrics for inaccuracy and conflicting descriptions of each format.

3.1 Endangerment Computation Model

The rule-based system uses a fuzzy model to estimate the endangerment level (i.e. high vs. middle vs. low) for the analysed file formats. The computation of the overall endangerment level is performed by integrating the view of the expert community (see Figure 2) and by using the associated fuzzy rule model (see Figure 3). The Endangerment

Computation Model (ECM) can be customized to model the policies of a particular organisation.

The model proposed for evaluating the endangerment level comprises three blocks of rules grouped by their impact level (see Figure 3). Each of the factors taken in account are evaluated based on the associated metrics. The analysis of risk factor calculations delivers three fold results. An "endangerment" output estimates the endangerment levels. A "conflicts" output analyses the conflicting information received from different sources. This analysis takes in account format properties that include: description, software count, vendor count, compression, versions count, existence period, complexity, dissemination, deprecation, genre, homepage, standard, migration, digital rights, popularity, web browser support, MIME, timestamp, etc. This module estimates the severity of the conflicts and their occuring rate. For example see Table 3 in detailed report section. Finally, we have defined the "inaccuracies" part that tracks inaccuracies associated with a particular file format, it estimates their severity level and their count. By combining the outputs of these three modules, the inference engine concludes the overall endangerment level and evaluates the risks for the analysed format. More about the risk factors is described in Section 4.3.

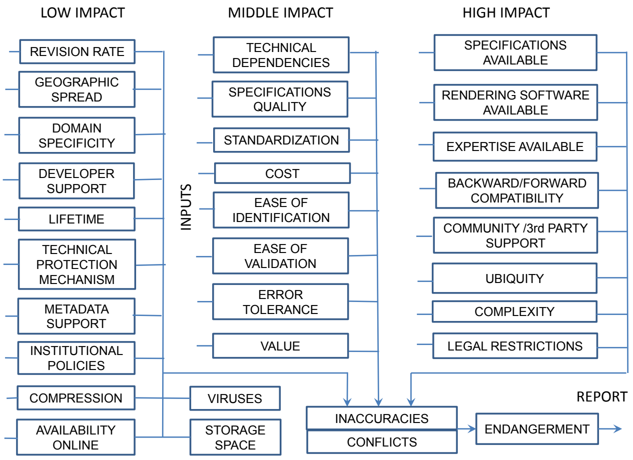


Figure 3: An inference model for calculation of endangerment level.

3.2 Metric Computation Model

The metrics for the rule "Complexity" in Figure 4 have different ranges for input values that are presented in angular braces. These ranges can be numerical, boolean or textual. The input values for these ranges can be retrieved from LOD repositories employing FFMA tool. As a sample for this rule we will analyze the PDF format. The metric "DISCLOSURE" becomes input value "yes" since it is an open standard ISO 32000 as stated in "Adobe" vendor documentation pointed by Fileinfo registry. This format is broadly used by thousands of vendors worldwide. The estimation of document numbers is hard to define because of different types of documentation like books, textual documents and HTML tutorials. We have counted 1662 tutorial documents and each of them has in average 2 pages. Number of formulas in documentation has low relevance in our opinion but it would make sense to estimate number of code snippets or screenshots. In that sense we counted this metric with 4 per page in average. Features count can be also found in documentation and is given by at least 10 top features but that can't be automated. We have found 8 color spaces. The effort to sustain information objects can be very different depending on organisation goals and can be measured in money amount and/or working hours. The intelligibility and understandability of this format is high since it can incorporate another formats, renders on different operation systems and has a high level of community and vendor support. PDF is supported by 28 software tools (see Table 2) that has middle level in our classification. As a part of training we found 10 test scenarios. PDF supports text, drawings, videos, audio, 3D maps, full-color graphics, photos and business logic. Rules of the format are very difficult to estimate since rule definition is vague. We found 19 rules meaning different aspects of the standard.

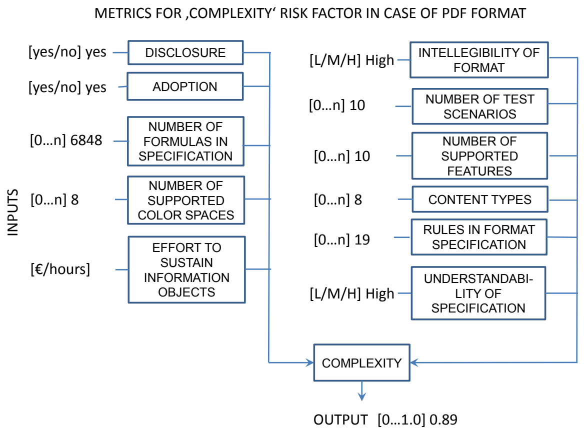

METRICS FOR ‚COMPLEXITY' RISK FACTOR IN CASE OF PDF FORMAT
INPUTS
[yes/no] yes
DISCLOSURE
[yes/no] yes
ADOPTION
[0...n] 6848
NUMBER OF FORMULAS IN SPECIFICATION
[0...n] 8
NUMBER OF SUPPORTED COLOR SPACES
[€/hours]
EFFORT TO SUSTAIN INFORMATION OBJECTS
[L/M/H] High
INTELLEGIBILITY OF FORMAT
[0...n] 10
NUMBER OF TEST SCENARIOS
[0...n] 10
NUMBER OF SUPPORTED FEATURES
[0...n] 8
CONTENT TYPES
[0...n] 19
RULES IN FORMAT SPECIFICATION
[L/M/H] High
UNDERSTANDABI-LITY OF SPECIFICATION
COMPLEXITY
OUTPUT [0...1.0] 0.89

Figure 4: An inference system for calculation of the complexity risk factor by employing of the associated metrics for the given file format.

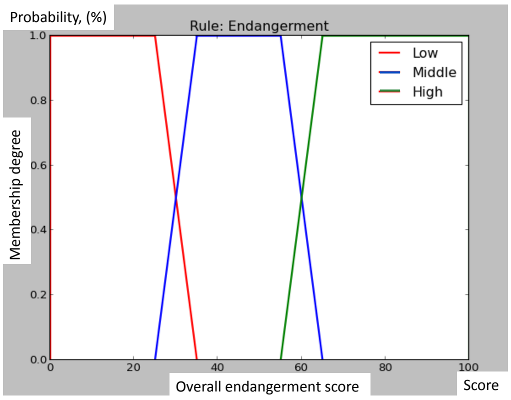
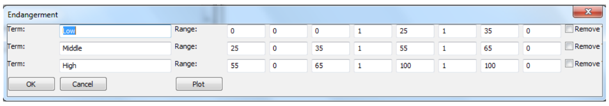


Figure 5: Plot of resulting endangerment level estimation as a result of all factors calculated by associated metrics.

The Figure 5 depicts graphical representation of previously



Term:	Low	Range:	0	0	0	1	25	1	35	0	□ Remove
Term:	Middle	Range:	25	0	35	1	55	1	65	0	□ Remove
Term:	High	Range:	55	0	65	1	100	1	100	0	□ Remove

Figure 6: Example fuzzy rule definition for endangerment rule.

defined fuzzy rules and their membership functions.

The Figure 6 shows an example fuzzy rule with associated values. These example demonstrates membership function $m(x)$ definition.

Using a fuzzy model allows us to deduce approximations of solid data points by aggregating multiple natural language data sources with varying levels of accuracy. The fuzzyfication is required in order to estimate format endangerment according to various facets of risk factors. Using fuzzification we obtain individual metrics for various risk factors. The fuzzyfication maps the numerical values to the decision variables by using the membership functions. By combining all defined fuzzyfied variables we can construct a hierarchical fuzzy inference system, since the output of a fuzzy inference module can be used as input for the next level of inference within the system. For example, the inference module for the complexity risk factor depicted in Figure 4 is used as input for the inference model presented in Figure 3.

A concrete example of complexity calculation is presented in Section 4. presented in the following sections. A fuzzy set estimates the risk level of a factor as belonging to the impact categories “Low”, “Middle” and “High”. This is decided by using membership functions as the ones presented within the Equations 1-5.

$$(U, m) = \{\frac{m(x_{LOW})}{x_{LOW}}, \frac{m(x_{MID})}{x_{MID}}, \frac{m(x_{HIGH})}{x_{HIGH}}\}, \quad (1)$$

$$x \in U \quad (2)$$

$$m(x_{LOW}) = \begin{cases} 1, & \text{if } 0 < x \leq 25, \\ -\frac{x}{10} + 3.5, & \text{if } 25 < x \leq 35, \end{cases} \quad (3)$$

$$m(x_{MID}) = \begin{cases} \frac{x}{10} - 2.5, & \text{if } 25 < x \leq 35, \\ 1, & \text{if } 35 < x \leq 55, \\ -\frac{x}{10} + 6.5, & \text{if } 55 < x \leq 65, \end{cases} \quad (4)$$

$$m(x_{HIGH}) = \begin{cases} \frac{x}{10} - 6.5, & \text{if } 55 < x \leq 65, \\ 1, & \text{if } 65 < x \leq 100. \end{cases} \quad (5)$$

Where (U, m) denotes a fuzzy set U with membership function $m(x)$. The concrete instances x belong to the set U with different degrees of membership quantified in numeric values - from not included ($m(x) = 0$) to fully included ($m(x) = 1$).

3.3 Knowledge Aggregation

The FFMA module[Graf and Gordea 2013] for aggregation of file format descriptions collects information from LOD repositories and enhances it by aggregation of expert knowledge. A specific exploitation context may customize which LOD repositories should be used and which file format properties are of interest for particular institutional context. The File Format Data Aggregation module is responsible for collecting descriptions on file format-related information from the open knowledge bases, while the FFMA engine combines the outcome of the module with the knowledge manually provided by domain experts. The acquired domain knowledge in stored in a local database and further used for reasoning in risk computation process. The external knowledge sources like DBPedia and Freebase manage huge amounts of LOD triples, which allows one to extract fragmental descriptions on file formats, software applications and software vendors.

4. EXPERIMENTAL EVALUATION

The goal of evaluation of format risks was the enhancement of FFMA knowledge base and validation of aggregated data. This process is described in the correctness calculation workflow (see Figure 2). Our hypothesis is that file format data automatically aggregated from LOD repositories will provide the fuzzy inference engine with valuable information and will enable correctness estimation for different file formats. The “high” confidence marked formats should indicate the currently most reliable file formats for digital preservation workflows. A Web service was developed that automatically retrieves file format related data from LOD repositories and performs reasoning on collected information employing specified risk factors. The collected information is processed, normalized, integrated into the knowledge base. The programming interface of this service supports querying for descriptions of the file formats, software, vendors and associated information. Service supports checking of availability of the information in the service database and retrieving data from LOD repositories if necessary. Another goal of our evaluation is the need to recognise that format is becoming obsolete and prepare adequate preservation planning, strategies and actions in response. Our approach should give an organisation a basis at hand that helps to choose a particular format and renderer. This decision should be the best choice for the organisation’s preservation programme. The employment of Fuzzy technique in comparison to FFMA[Graf and Gordea 2013] approach is more flexible and emulates a human expert by concept of partial truth, whereas FFMA risk system knows only True/False modes of truth.

4.1 Evaluation Data Set

For evaluation purposes a subset of 13 representative, well known file formats was selected. The *GIF*, *PNG*, *JPG*, *BMP* and *TIF* formats belong to the raster graphics genre. *MP3* is the most used audio format, while the *PDF* format is mostly used for document formats, having multiple versions and being well supported by Adobe Acrobat toolset. The *HTML* format also has multiple versions and is used for the creation of Web pages. The *DOC* and *PPT* are Microsoft formats supporting creation of multimedia documents and presentations. Some outdated file formats are represented by *MAC*, *SXW* and *DXF*. The *MAC* is a bitmap graphic format for the Macintosh, one of the first painting programs for this OS, supporting greyscale-only graphics. The *SXW* is an outdated text format for OpenOffice, while *DXF* is a vector graphic format for AutoCAD.

4.2 Computation of Risk Factors

The previously defined rules should be organized in order to process input values and to infer appropriate conclusions. As an example, the rule-base system may start endangerment identification for PDF format with the inference engine of the "Complexity" factor in Figure 4 which comprises 11 fuzzy preconditions. The particular input values are depicted by the rectangles sorted by impact level that was evaluated from the survey. Having input values on the left side and running calculations we receive a confidence level value 0.89 on the output. According to our FLC definitions depicted in Figure 5 that means that resulting confidence level is "high". The value "high" is a result of matching the numerical output value 0.89 to the fuzzy rule for calculation of confidence level using member functions in Equation 1, where "low" is defined for values in range from 0 to 0.35, "middle" from 0.25 to 0.65 and "high" from 0.55 to 1.0 respectively. Therefore, the input value of the "Complexity" factor in Figure 3 is 0.89. The Expert System calculates the complexity level of the format as "high" if most of the metrics after fuzzification produce total output value greater than 0.67. Each of the metrics can again be formulated as a fuzzy rule according to preferences of particular institution. Fuzzifying this value we map it to the associated numerical value using FLC input variables definition. Aggregating all rule outputs we defuzzify the output value of the total endangerment level that is "high" and map it to the resulting number 0.93.

An input variable "Resulting Risk" contains three membership functions flagged by the linguistic variables "Low, Middle and High". A corresponding graphical representation is shown in Figure 5. The values for these linguistic variables range from 0 to 1 and are coming from the inference engine. For simplicity we transform these values to percents. Therefore, format risk can be defined as high if its value matches in a range between 55 and 100 percent. In contrast middle risk values are between 25 and 65 percent. Finally values between 0 and 35 percent indicate that there is low risk for analyzed file format.

Table 1 shows an adapted set of file format risk factor rating results from a file format study conducted by Heather Ryan[Ryan 2014]. The study was conducted among 11 digital preservation experts over three rounds. The relevance of particular factor as an indicator of file format endangerment, from the left column on file format risk is defined by values from 1 to 3. Value 3 in this table stands for "Very relevant", 2 for "Somewhat relevant"and 1 for "Not relevant at all" respectively. The most relevant factors according to evaluation are listed first. The column "SUM" depicts the sum of all votes. The average relevance per factor was calculated and depicted in the "AVG" column. Also the total endangerment value for each factor wascalculated and presented in the column "Endangerment level". This row demonstrates how relevant the factor is for the whole format estimation by associated linguistic values in range between "Middle" and "High". The detailed information about the spread of the distribution of the various expert views is presented in risk factor analysis[Ryan 2014]. This should provide information about the degree to which the experts agreed or not regarding particular risk factors.

The suggested factors cover most of the risk factors identified in FFMA. Merging these two sets we get a basis for fuzzy system. The main conclusion from the review presented in Tables 1 and 2 is that there is a need for some metrics describing file formats. Such metrics can be automatically provided by the extended FFMA risk model[Graf and Gordea 2013]. By metrics definition we will stick by previously presented in FFMA and in survey simple range Low/Middle/High. The goal by defining metrics is to automate an evaluation of file format risk. In some situations many metrics probably are not realistic since no universal standards for them exist but nevertheless automation can be possible for institutional use cases with good documented workflows. Estimation of risk factor risks is impossible without definition of quality metrics and relevant semantics.

4.3 Risk Factors with High Impact

The description of the high impact risk factors is presented below. The more detailed description and analysis is presented in the file format study of Heather Ryan[Ryan 2014]

- The 'Backward/Forward Compatibility' factor influences how easily and inexpensively content in original format can be accessed, migrated and meaningfully rendered and is a mitigating factor of endangerment or obsolescence of a file format. Measuring of this factor employs information about software that fails in reading an older format, about font substitution failures and about automatically adjusting the color space. Another attributes for this factor are well documented format specification, rendering software number and documentation, licence management, number of versions, release notes and direct testing support measurement of backward compatibility that should be verified by a human.
- The 'Community/3rd Party Support' factor enables people to implement the format through the existence of multiple independent implementations using the same format. This ensures that the format is stable and well-defined. It can be measured by number of communities, by number of software applications supporting it, by trends of software support compared to previous time period, by emulation environments and by counting the number of users or files. It is possible, proprietary formats are more difficult to be supported by a community. This factor depends on how much of the specifications are published and if a file format contains patented parts or techniques.
- The 'Complexity' factor can have a different meaning for different institutions. For example, the level of complexity for PDF is so high that the costs of providing access might become unsustainable. Measurement of complexity requires accurate generation of a representation network, which is difficult to automate. It is dependent on specifications quality, implementations number for the same functionality within a document, number of testing scenarios. Optionally supported features complicate the evaluation of compatibility. The feature rich specification such as JPEG2000 is more complex than a very simple specification such as that of a GIF file. In a long term preservation strategy it can be much harder to migrate or continue rendering a

Table 1: Risk factors rating for digital preservation of file formats from the survey

Risk Factor	SUM	AVG	Experts Number	Endangerment Level
Specifications Available	33	3.000	11	high
Rendering Software Available	32	2.909	11	high
Expertise Available	30	2.727	11	high
Backward/Forward Compatibility	29	2.636	11	high
Community/3rd Party Support	29	2.636	11	high
Ubiquity	29	2.636	11	high
Complexity	27	2.455	11	high
Legal Restrictions	27	2.455	11	high
Technical Dependencies	26	2.364	11	middle
Specification Quality	23	2.300	10	middle
Standardization	25	2.273	11	middle
Cost	25	2.273	11	middle
Ease of Identification	24	2.182	11	middle
Ease of Validation	24	2.182	11	middle
Error-tolerance	22	2.091	11	middle
Value	20	2.000	10	middle
Revision Rate	21	1.909	11	low
Geographic Spread	19	1.900	10	low
Domain Specificity	19	1.900	10	low
Developer/Corporate Support	20	1.818	11	low
Lifetime	20	1.818	11	low
Technical Protection Mechanism	20	1.818	11	low
Metadata Support	18	1.636	11	low
Institutional Policies	16	1.600	10	low
Compression	17	1.545	11	low
Availability Online	15	1.500	10	low
Storage Space	15	1.364	11	low
Viruses	13	1.300	10	low

highly complex file format. Complexity attributes are depicted in Figure 4.

- The factor 'Expertise Available' impacts the long-term viability of rendering, migration or emulation. A digital preservation expert needs to understand the whole platform especially proprietary formats. The attributes for expertise estimation are expert skill level, experience, software documentation and its date, communities available and its size, age of technology, popularity of technology.

- The factor 'Legal Restrictions' handles restrictions caused by licensing, which can be a barrier to software developers providing support for the format. This can be problematic when selecting an emulation strategy for long term preservation. The PREMIS metadata standard has semantic units for capturing this, that might need to be extended. The EU project 'KEEP' has many case studies on this topic. This factor is dependent on licence and number of patents.

- The factor 'Rendering Software Available' is important for understanding when renderability is compromised and then institute the appropriate preservation planning, strategies and actions necessary to ensure it. This factor can be evaluated by testing, licencing, contacting vendors, using characterisation software and technology watch.

- The factor 'Ubiquity' is based on the assumption is that widely used format will be less likely subject to obsolescence. This depends on things like the viability of the supplier, whether it is proprietary or not and the emergence of new more interesting formats. Well used file formats have both active user communities and are more attractive to commercial companies to provide new products to support old formats. The more ubiquitous a file format, the wider the availability of toolsets for rendering, validation, identification, migration and emulation. Ubiquity attributes are market survey research, popularity, vendor information, proprietary-ness, number of files, web search, and number of software implementations.

- The factor 'Specification Quality' expresses the expectation that a specification be complete and well written. The better the specification, the better any new implementation will be. As OAIS notes, sometimes source code for a renderer is itself representation information for a format. It is dependent on levels of satisfaction and specification.

An overview of the computed low level risks for the formats included in the evaluation set is presented in Table 2. The values and the interpretations of the most important 23 risk factors are presented. Within this representation, the "+" sign stands for *true* while the "-" sign means *false*. L depicts low risk, M means middle risk and H stands for high risk. This table shows that among evaluated formats, the DOC format has the highest number of supported software, whereas for SXW only one software tool was documented in LOD repositories. The remaining formats have different software numbers, mostly between 10 and 40.

The different risk scores for DOC (low) and PPT (middle) could be explained with larger amount on software tools automatically detected for DOC (164) comparing to four for PPT and also with more descriptions for DOC format. Additionally, for DOC the genre, creation date, publisher and creator information were retrieved, whereas these factors are missing for *PPT.* This does not mean that such information does not exist for PPT, it only indicates that this is not included or not found in LOD repositories. The same consideration is valid for the "software count" value 12 of *MP3* format. It is known that there should be much more associated software tools that are able to handle this format.

At this point it should be stated that not all formats were analyzed and that evaluated results currently require verification by human experts and further optimisation of calculation methods. Evaluation results presented in Table 2

Table 2: Exemplarily selected file formats with retrieved information for associated measurement metrics

Risk Factor	GIF	PNG	MP3	PDF	JPG	DOC	HTML	TIF	BMP	PPT	MAC	SXW	DXF
Is Popular Format	5/L	5/L	5/L	5/L	5/L	5/L	5/L	5/L	5/L	5/L	2/H	3/M	5/L
Operation Systems	3/M	4/L	3/M	6/L	4/L	5/L	4/L	3/L	2/M	5/L	2/M	3/M	4/M
Software Count	18/M	21/M	14/M	28/M	17/M	164/L	39/L	135/L	18/M	15/M	122/L	1/H	21/M
Vendors Count	3/L	1/M	3/L	2/L	1/M	1/M	1/M	1/M	1/M	1/M	1/M	1/M	1/M
Versions Count	2/M	3/M	1/L	17/H	9/H	15/H	7/H	9/H	7/H	7/H	1/L	1/L	23/H
Has Description	3/M	3/M	2/H	3/M	2/H	3/M	2/H	3/M	2/H	2/H	2/H	2/H	2/H
Has MIME type	+/L	+/L	+/L	+/L	+/L	+/L	+/L	+/L	+/L	+/L	-/H	-/H	-/H
Existence Period	+/L	+/L	+/L	+/L	+/L	+/L	+/L	+/L	+/L	+/L	+/L	+/L	+/L
Is Complex Format	-/L	-/L	-/L	+/H	-/L	-/L	+/H	+/H	-/L	-/L	-/L	+/H	+/H
Is Wide Disseminated	+/L	+/L	+/L	+/L	+/L	+/L	+/L	+/L	+/L	+/L	-/H	-/H	-/H
Is Outdated or Deprecated	-/L	-/L	-/L	-/L	-/L	+/H	+/H	-/L	-/L	+/H	+/H	+/H	+/H
Has Genre	+/L	+/L	+/L	+/L	+/L	+/L	+/L	+/L	-/H	-/H	-/H	-/H	-/H
Has Homepage	+/L	-/H	-/H	+/L	-/H	-/H	-/H	-/H	+/L	-/H	-/H	-/H	-/H
Is Open (Standardised)	+/L	+/L	+/L	+/L	+/L	-/H	+/L	-/H	-/H	-/H	-/H	-/H	-/H
Has Creation Date	+/L	+/L	+/L	+/L	-/H	+/L	+/L	+/L	-/H	-/H	-/H	-/H	-/H
Has File Migration Support	+/L	+/L	+/L	+/L	+/L	+/L	+/L	+/L	+/L	+/L	+/L	+/L	+/L
Digital Rights Information	-/H	-/H	-/H	-/H	-/H	-/H	-/H	-/H	-/H	-/H	-/H	-/H	-/H
Has Publisher Information	+/L	-/H	+/L	+/L	+/L	+/L	+/L	-/H	+/L	-/H	-/H	-/H	-/H
Has Creator Information	+/L	-/H	+/L	+/L	+/L	+/L	+/L	-/H	+/L	-/H	-/H	-/H	-/H
Has Compression Support	-/L	-/L	-/L	-/L	-/L	-/L	-/L	+/H	-/L	-/L	-/L	-/L	-/L
Supported by Web Browser	+/L	+/L	+/L	+/L	+/L	+/L	+/L	+/L	+/L	+/L	+/L	+/L	+/L
Has Vendor Support	+/L	+/L	+/L	+/L	+/L	+/L	+/L	+/L	+/L	+/L	+/L	+/L	+/L

Table 3: Exemplarily selected file formats with retrieved correctness information

Format	Expert Knowledge	Inaccuracies	Conflicts	Confidence Level
PDF	High	2	3	Middle
JP2	High	3	6	Low
JPG	Middle	1	1	High
JPX	Low	1	1	High
PNG	Middle	1	1	High
GIF	Middle	1	2	Middle
DOCX	Low	0	1	High
TIFF	High	0	1	High

are limited to the information automatically collected from LOD repositories mentioned above, and are customized by the applied expert rules. Therefore these results cannot be regarded as absolutely accurate, but they provide a good overview of the possible preservation risks related to the given file formats. The classification settings for risk factors are institutionally dependent and is a matter of discussion and a future work. The default thresholds are defined based on the accessible expert knowledge and could be customized according to preferences of particular user.

4.4 Detailed Report

The evaluation demonstrates 3 that the given approach shares expertise and supports contradiction comparison for one institution and addresses specific risks within file formats. Information support provided by the Expert System helps in solving practical digital preservation issues. But in order to generate higher value in aggregating the data sources and exposing conflicts and inaccuracies this tool needs more and better quality data sources. The column "Inaccuracies" shows the number of wrong or inaccurate automaticly retrieved statements detected by experts. The column "Conflicts" demonstrates the number of controversial automatically retrieved statements detected by experts.

Although FFMA provides valuable information that well describes the evaluated formats, the accuracy of data collected in the FFMA knowledge base should be examined by experts. The PDF is marked as a non-compressed format, but experts state that PDF nearly always uses flat compression, whereas a whole array of compression methods may be used for images. PNG, JPG and GIF are flagged in FFMA as uncompressed whereas they have compression. The Jpeg2000 format according to FFMA is not supported by any software and does not have a MIME type, is frequently used and is supported by web browsers. In reality these factors are wrong in FFMA. The JPX format is marked as a non-compressed that should be less complex than JP2, but actually it is an extension of Jpeg2000 with added complexity. The GIF is marked as having the highest risk. The TIFF format should have higher risk than PDF or DOCX. The PDF can be a container for Jpeg2000 which is considered high-risk in FFMA. The mentioned confidence levels should not be regarded as a preservation risk estimation for associated format. Currently FFMA provides generalized information about formats, without addressing specific risks within formats.It should be mentioned that presented confidence levels are considered in relation of FFMA results to expert knowledge. These are FFMA evaluation results and should help the user to resolve these contradictions.

5. CONCLUSIONS

In this work we presented an approach for bringing together information automatically aggregated from open sources and an expert knowledge related to digital preservation. The main contribution of this work is the definition and computation of fuzzy logic for metrics generation in order to support digital preservation experts in semi-automatic estimation of "institutional obsolescence" for file formats. We aggregated a solid knowledge base from linked open data repositories. In the correctness report we exposed conflicts and inaccuracies in these data in order to improve the quality of a risk analysis in the digital preservation domain. This method facilitates decision making with regard to the preservation of digital content in libraries and archives using expert knowledge as a basis. We have developed a tool for aggregating file format descriptions that exploits available linked data resources and uses expert models to infer knowledge regarding the long-term preservation of digital content. The ontology mapping technique that comprises expert rules and clustering is employed for collecting the information from the web and integrating it in a common representation.

We employed fuzzy logic techniques for processing aggregated information about formats using metrics in order to bring conflicted and incorrect information to the surface for correction and improvement by the community. The analysis of a sub-set of results from a study on the risk factors for file formats was integrated in a fuzzy model and is presented in the evaluation section.

The evaluation demonstrates that the given approach shares expertise and supports contradiction comparison for one institution and addresses specific risks within file formats. Information support provided by the Expert System helps in solving practical digital preservation issues. But in order to generate higher value in aggregating the data sources and exposing conflicts and inaccuracies this tool needs more and better quality data sources. The analysis and measurement provided by developed Expert System is about the reduction of uncertainty and not about the elimination of it. Using our system with its metrics we have the ability to measure and the ability to think about how we can use these measurements.

As future work we plan to increase the amount of aggregated information, to extend an Expert System with additional fuzzy rules and to improve its accuracy and quality of the outputs.

6. ACKNOWLEDGMENTS

This work was partially supported by the EU FP7 Project SCAPE (GA#270137) `www.scape-project.eu`. The authors wish to thank Paul Wheatley from the British Library for his thoughts on the topic.

7. REFERENCES

[Ayris et al. 2008] P. Ayris, R. Davies, R. McLeod, R. Miao, H. Shenton, and P. Wheatley. 2008. *The LIFE2 final project report.* Final project report. LIFE Project, London, UK.

[David Tarrant 2011] Leslie Carr David Tarrant, Steve Hitchcock. 2011. Where the Semantic Web and Web 2.0 Meet Format Risk Management: P2 Registry. *International Journal of Digital Curation* 6, 1 (2011), 165–182.

[Gordea et al. 2011] Sergiu Gordea, Andrew Lindley, and Roman Graf. 2011. Computing Recommendations for Long Term Data Accessibility basing on Open Knowledge and Linked Data. *Joint proceedings of the RecSys 2011 Workshops Decisions@RecSys'11 and UCERSTI 2* 811 (November 2011), 51–58.

[Graf and Gordea 2012] Roman Graf and Sergiu Gordea. 2012. Aggregating a Knowledge Base of File Formats from Linked Open Data. *Proceedings of the 9th International Conference on Preservation of Digital Objects* poster (October 2012), 292–293.

[Graf and Gordea 2013] Roman Graf and Sergiu Gordea. 2013. A Risk Analysis of File Formats for Preservation Planning. In *Proceedings of the 10th International Conference on Preservation of Digital Objects (iPres2013).* Biblioteca Nacional de Portugal, Lisboa, Lissabon, Portugal, 177–186.

[Hunter and Choudhury 2006] J. Hunter and S. Choudhury. 2006. PANIC: an integrated approach to the preservation of composite digital objects using Semantic Web services. *International Journal on Digital Libraries* 6, (2) (September 2006), 174–183.

[Jackson 2012] Andrew N. Jackson. 2012. Formats over Time: Exploring UK Web History. *Proceedings of the 9th International Conference on Preservation of Digital Objects* (October 2012), 155–158.

[Lee 1990] C.-C. Lee. 1990. Fuzzy logic in control systems: fuzzy logic controller. I. *Systems, Man and Cybernetics, IEEE Transactions on* 20, 2 (1990), 404–418. `DOI:http://dx.doi.org/10.1109/21.52551`

[Pearson and Webb 2008] David Pearson and Colin Webb. 2008. Defining File Format Obsolescence: A Risky Journey. *The International Journal of Digital Curation* Vol 3, No 1 (July 2008), 89–106.

[Rosenthal 2010] David S.H. Rosenthal. 2010. Format obsolescence: assessing the threat and the defenses. *Library Hi Tech* 28, 2 (2010), 195–210.

[Rothenberg 2012] Jeff Rothenberg. 2012. Digital Preservation in Perspective: How far have we come, and what's next? *Future Perfect 2012* (2012).

[Ryan 2013] Heather Ryan. 2013. File Format Study. *School of Information and Library Science, University of North Carolina at Chapel Hill* 2 (2013).

[Ryan 2014] Heather Ryan. 2014. Occam's Razor and File Format Endangerment Factors. In *Proceedings of the 11th International Conference on Preservation of Digital Objects (iPres2014) (accepted for publication).* Melbourne, Australia, 10.

[Sii et al. 2001] How Sing Sii, Tom Ruxton, and Jin Wang. 2001. A fuzzy-logic-based approach to qualitative safety modelling for marine systems. *Reliability Engineering & System Safety* 73, 1 (2001), 19 – 34. `DOI:http://dx.doi.org/10.1016/S0951-8320(01)00023-0`

[Vermaaten et al. 2012] Sally Vermaaten, Brian Lavoie, and Priscilla Caplan. 2012. Identifying Threats to Successful Digital Preservation: the SPOT Model Rsik Assessment. *D-Lib Magazine* 18, 9/10 (September 2012).

[Zadeh 1996] Lotfi A. Zadeh. 1996. Fuzzy logic = computing with words. *Fuzzy Systems, IEEE Transactions on* 4, 2 (1996), 103–111. `DOI:http://dx.doi.org/10.1109/91.493904`

VPlan - Ontology for Collection of Process Verification Data

Tomasz Miksa
SBA Research
Vienna, Austria
tmiksa@sba-research.org

Ricardo Vieira
José Barateiro
INESC-ID Information Systems Group
& LNEC
Lisbon, Portugal

Andreas Rauber
Vienna University of Technology
& SBA Research
Vienna, Austria
rauber@ifs.tuwien.ac.at

ABSTRACT

The reproducibility of modern research depends on the possibility to faithfully rerun the complex and distributed data transformation processes which were executed by scientists in order to make new scientific breakthroughs. New methods and frameworks try to address this problem by collecting evidence used for verification of such experiments. However, there is still a lack of a flexible data model which would address all of the needs of these methods. This paper presents the VPlan ontology designed for the purpose of organizing and storing of data collected for verification of preserved processes. The VPlan ontology stores and links the data extracted from the preserved process. Furthermore, it includes descriptions of actions taken to collect the data, as well as provides a clear break down of requirements that lead to its collection. We demonstrate the usage of the VPlan ontology within the preservation process and describe in detail its alignment with the Verification Framework (VFramework). In order to illustrate its applicability to the eScience domain, we evaluate it on a use case from the civil engineering domain, which is an example of a typical sensor data analysis process.

1. INTRODUCTION

The preservation of entire processes and workflows has already gained the interest of the digital preservation community [18]. There are a number of research projects [3, 11] addressing the challenges of keeping processes available in the long term. They deliver tools [8] and frameworks [17] which try to address the problem of not only preserving the data which is produced at the output of the eScience experiments, but also preserving the way the results were obtained. This includes preservation of complex and very often distributed processes which captured, processed, integrated or visualised the data. Despite these advances, the problem of reproducibility of modern data-intensive science remains unsolved and is currently receiving the attention of publishers [12], funding agencies [9] and researchers themselves [4]. As a result, scientists are often required to create data management plans in which they describe the data produced by their experiments. This solves the problem partially, because the information on processes used in the experiments are still not detailed enough. Process management plans [14] complement the data management plans with information on processes, but they are still not fully implemented.

Most of these efforts focus only on the problem of preserving the experimental data and documenting the processes executed to obtain these results. However, information needed for verification and validation of the redeployed process must also be captured. The verification of redeployed processes is a complex task and depends on many things: the way the processes are specified, the drivers for their preservation, the preservation strategies applied; the reasons for the redeployment, the redeployment environments, and so on. Such information must be collected at the time of process execution and is later used to prove that the process running in the redeployment environment is performing in the way it was originally meant. This may be crucial in litigation cases when the correctness of the original process executed at some time in the past could be questioned and the only way to check this is to re-run the original process. The verification can only be reliable when the requirements used for the verification are well structured and the processes of data capturing and redeployment quality metrics calculation are clearly defined.

In [13] we presented the VFramework which defines a framework for verification of preserved and redeployed processes. In this paper we present the VPlan which is an ontology for collection of process verification data. The VPlan stores the information collected during application of the VFramework. It integrates well with the TIMBUS Context Model [2, 11] and makes use of the ArchiMate [20] modelling language to describe the data capture processes. It also links the significant properties and metrics, which are used for verification, to the real location of data. In this paper we also demonstrate the applicability of the VPlan to the verification of preserved and redeployed eScience processes. We use a use case from the civil engineering domain which is an example of a typical sensor data analysis process.

The paper is organized as follows. Section 2 presents the state of the art. In Section 3 the VPlan is described and mapping to the VFramework is provided. Section 4 describes usage of the VPlan in the eScience use case. We provide conclusions and future work in Section 5.

2. STATE OF THE ART

This section discusses the most important work related to the verification and validation of preserved processes. We also place this work in the context of the TIMBUS Preservation Process and explain concepts that impacted the design of the VPlan.

2.1 Verification framework

In [6] a conceptual framework for evaluation of emulation results was presented. It was demonstrated in [5] that the framework can be successfully applied to evaluate the conformance and performance quality of applications and simple processes redeployed in an emulator. This was demonstrated in case studies in which the framework was used to evaluate the emulation of a video game and an accounting program. The VFramework presented in [13] is a refinement of that framework for complex, potentially distributed processes. It provides detailed specification of actions which have to be performed for verification of redeployed processes. The VFramework is presented in Figure 1 and consists of two sequences of actions. "The first one (depicted in blue) is performed in the original environment. The result obtained from the execution of each step is written into the VPlan. The second sequence (depicted in green) is performed in the redeployment environment. The necessary information for completion of each of the steps is read from the VPlan." [13] By original environment we mean a system in which the process is executed. The redeployment environment is the system to which the process will be moved when a decision to rerun the preserved process is taken. The redeployment can take place at any time in the future when the original platform is not available anymore. Hence, it may be necessary to re-engineer the process in order to fit it into a new system.

2.2 TIMBUS Preservation Process

In [18] the TIMBUS Preservation Process for preservation of processes is presented and applied to an eScience process. The authors explain three phases of the approach: plan, preserve and redeploy. The TIMBUS Preservation Process assumes that the verification data is collected during the preserve phase and used for verification of the process in the redeploy phase. The VFramework [13] provides a detailed list of steps for performing verification when executing the TIMBUS Preservation Process. The VPlan presented in this paper describes an ontology for collection of verification data. Detailed information on the TIMBUS Preservation Process can also be found in [21].

2.3 Process modelling

Processes, as organized sets of activities performed to achieve specific desired outcomes, are something that exists in all organizations and might be described and documented in many different ways. The description of a process using a set of key concepts and relations is typically known as process modelling. Modelling enables a common understanding easing the analysis of a process [1]. There are several techniques to model processes depending on the pretended analysis, such as flow charts, data flows, and role activity diagrams [1]. The most known and used technique and language to describe the flow of a business process is the Business Process Modelling Notation (BPMN) [16].

Enterprise Architecture (EA) is a coherent set of principles, methods and models to design, analyse, change and manage organizations through four main architecture domains: business, data, application and technology. However, in order to properly describe the main concepts of EA and the dependencies between domains, BPMN is insufficient [19]. Therefore EA languages emerged in order to address the existing gap. ArchiMate [7] represents the culmination of years of work in the area of EA modelling languages and frameworks and is one of the most used EA languages nowadays. It provides high-level abstract concepts divided into three tightly connected EA layers: the business layer, the application layer, and the technology layer. It is a mature language with extensive use and practice where elements and relationships are clearly defined and explained [19].

Taking into account the advantages of Archimate against the common process modelling languages, Archimate is used to model the required processes in the VPlan presented in this paper, namely the preserved process, the capture processes and, if they exist, the determinism transformation processes.

2.4 Ontologies

Provenance ontologies seem a natural candidate to be used at least as a basis for extension in order to address the requirements of the VFramework. The Open Provenance Model[1] has a corresponding OPMO[2] ontology. It describes process execution, but does not allow for definition of one's own metrics. Similarly the information contained in the Janus [15] ontology describes execution of a workflow, i.e. data exchanged between workflow elements, timestamps, and so on. This information is useful for modelling of the process instance execution, but does not provide information on the significant properties, metrics or conditions in which the capturing took place. The Wf4Ever[3] project uses the wfprov[4] ontology that is capable of storing information about the execution and the parameters of a workflow, but there is also no information on significant properties or capture processes. Furthermore, both Janus and wfprov are limited to formally specified processes like workflows. Achieving the functionality of the VPlan by linking any other ontology to the OPMO, wfprov or Janus ontologies would not be possible and may lead to semantic inconsistencies between the concepts. None of the existing ontologies is suitable to fully address the requirements of the VFramework and neither is the composition of them.

3. VPLAN

The VPlan is an ontology-based document for storing and organizing information collected during the VFramework application. The following subsections describe: its structure, integration with the Context Model and mapping to the VFramework steps.

[1] http://openprovenance.org/

[2] http://openprovenance.org/model/opmo

[3] http://www.wf4ever-project.org/

[4] http://purl.org/wf4ever/wfprov

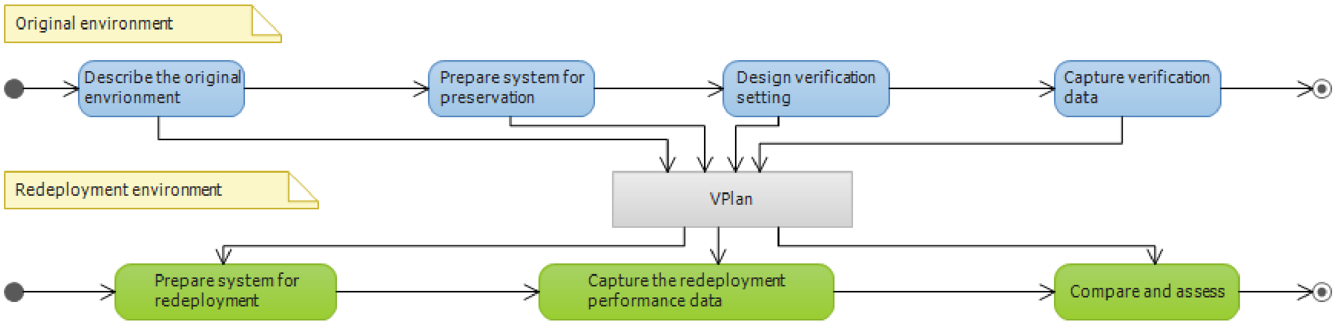

Original environment
Describe the original envrionment
Prepare system for preservation
Design verification setting
Capture verification data
VPlan
Redeployment environment
Prepare system for redeployment
Capture the redeployment performance data
Compare and assess

Figure 1: VFramework [13].

3.1 Overview

The VPlan[5] is created when the original process is preserved. It is accessed during the redeployment phase. The VPlan is created per process and it contains process instances which can verify particular process execution.

Figure 2 depicts the concept map of the VPlan. The names of the concepts correspond to the concepts defined in [13]. The light blue boxes are the classes, e.g. *VPlan*, *Metric*, *RedeploymentScenario*, and so on. The named arrows connecting the light blue boxes are object properties, e.g. *measures*, *appliesToScenario*, *hasInstance*, and so on. The arrows that point to the green boxes are the data properties, namely: *isLocatedAt*, *hasTextDescription* and *isInline*. There are also five dark blue boxes, which are individuals used for creating an enumeration for the *MetricTargetOperator* class. Finally, there are 3 grey boxes which depict elements imported to the VPlan by importing the TIMBUS Context Model.

In general the VPlan links the requirements expressed by significant properties and metrics with the way they are measured. To describe the measurement process, the information on process instances and capturing processes is provided. The VPlan uses the Context Model to precisely depict from which process' part the information was captured. Moreover, it includes capturing processes, which were originally modelled in ArchiMate and later converted to an ontology in order to document the way the data was collected. Finally, the VPlan stores not only information on data location used to run the process (process instances), but also the data which was captured from the process for calculation of metrics.

3.2 Relation to the Context Model

Due to the fact that the VPlan is an OWL[6] document, it benefits from integration with other ontologies. By default it is integrated with the TIMBUS Context Model. Furthermore, if different concepts are needed, the VPlan can integrate with any other existing ontology. The VPlan uses the Context Model in four different ways:

- import of the Context Model concepts at the model level,
- import of the preserved process at the instance level,
- import of the capture process at the instance level,
- import of the determinism transformation process at the instance level.

Figure 3 illustrates the relation of the VPlan to the Context Model. Each of the cases is discussed in the next subsections.

3.2.1 *Import of the Context Model at the model level*

The VPlan is coupled with the Context Model at the model level. This is one of the fundamental assumptions. Due to this coupling, the VPlan can make an extensive use of the machine-readable representation of the process. Moreover, the Context Model is based on the ArchiMate specification which is a recognized standard by many Enterprise Architects. Therefore, reuse of concepts from the Context Model (and indirectly from the ArichMate) in the VPlan facilitates VPlan understanding to users from these communities.

3.2.2 *Import of preserved process at the instance level*

The TIMBUS preservation framework assumes that in one of the initial steps a Context Model of the preserved process is created. Because the VPlan is always targeted at a particular process, then a coupling of the VPlan and the Context Model of the preserved process is natural. This is achieved by importing the ontology-based representation of the process into the instance of the VPlan. As a result, the redeployment scenarios, measurement points and levels of comparison (see [13] for definitions explanation) can easily be specified.

The redeployment scenarios can be described by connecting the *RedeploymentScenario* individual with each process step of the preserved process. As a consequence, further dependencies of each process's step can be inferred automatically without the need for explicit specification. When it comes to the specification of measurement points, they can be pointed directly to the preserved process and thus any ambiguities, which could stem from a verbal description, are removed. The levels of comparison are implicit and depend on the kind of process element to which the measurement point links.

3.2.3 *Import of capture processes at the instance level*

The VPlan requires that for each of the metrics a capture process is defined which describes how the data, which is

[5] http://timbus.teco.edu/svn/public/ontologies/VPlan.owl
[6] http://www.w3.org/TR/owl2-overview/

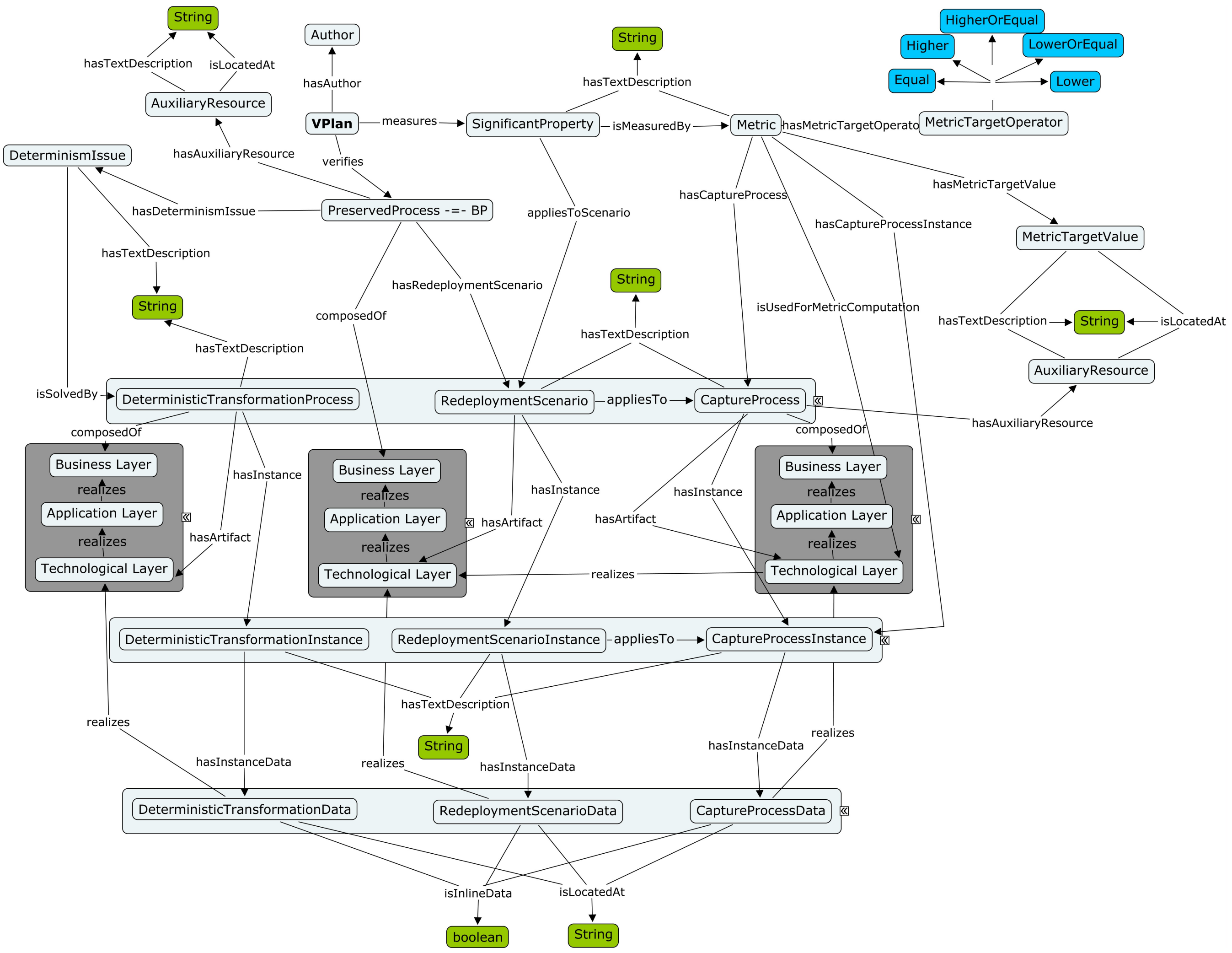


Figure 2: VPlan.

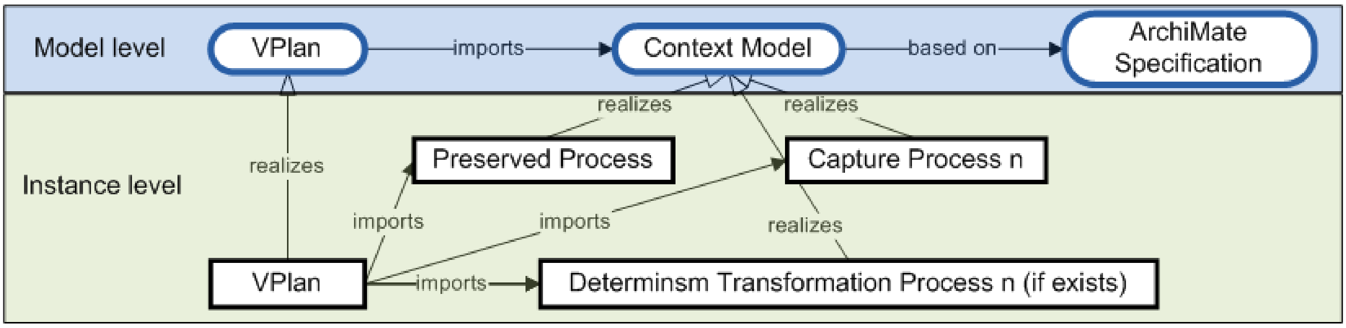


Figure 3: Differentiation between the VPlan model and the instance and an overview of imports made to the VPlan.

later used for metric computation, is extracted from the process. A similar approach was taken to the one from the Section 3.2.2 regarding the import of the preserved process model. Thus, each capture process is first modelled in ArchiMate, then converted to the ontology and finally imported to the VPlan.

Import of the capture process into the VPlan allows linking of the elements of the capture process with the elements of the preserved process. The link is essential, because in this way the generic process of capturing becomes concrete for the given preserved process. In other words, this link specifies the measurement point. For example, most of the capture processes provide at their output a file with some data extracted from the process. In order to state from which part of the process and at which component the capturing took place, a link between the *CaptureProcess* and the *PreservedProcess* is established.

3.2.4 Import of determinism transformation processes at the instance level

When the process is not deterministic during its execution, i.e. has different characteristic and outputs for the same input data, then it is impossible to conduct faithful verification. The VFramework foresees such a situation and assumes that for the purpose of verification the process part which introduces the lack of determinism can be removed or substituted with a deterministic one. Due to this fact, the VPlan holds information on determinism transformation processes. These processes describe what has to be done in order to make the preserved process deterministic for the purpose of verification. Similar to the capture processes described in the section above, the determinism transformation processes are modelled in ArchiMate, using the Archi[7] tool, converted to ontology and then imported to the VPlan.

3.3 Mapping to the VFramework

In this section the mapping of the VFramework steps to the VPlan classes is presented. The aim of the mapping is to demonstrate, that the VPlan fulfils the requirements of the VFramework. For this reason, two figures depicting mapping of concepts in the original and in the redeployment environment were created and are discussed in the consecutive subsections.

3.3.1 Original environment

The VFramework steps that are executed in the original environment focus on collection of process information. At this phase the VPlan is created and filled with data. The Figure 4 depicts which VPlan classes are used at which step of the VFramework application. The numbers on the arrows depict the concrete steps and substeps of the VFramework. If all substeps of a given step of the VFramework are making use of a given class, then only a number of a step is provided on the arrow, e.g. *AuxiliaryResource* is used at all of the substeps of the ”Describe the original environment” step of the VFramework, hence only 1 is used instead of 1.1/2/3/4.

In the first step of the VFramework, which is ”Describe the original environment”, not only the process and its context is described, but also the redeployment scenarios, verification instances and significant properties. According to the Figure 4 all these concepts are mapped to the respective classes.

In the second step of the VFramework, which is ”Prepare system for preservation”, a precise analysis of the process and its dependencies is conducted. This is the moment when the Context Model of the process is needed. The internal and external interactions of the process which are identified are modelled in the Context Model. The process boundaries are defined using *RedeploymentScenario* by specifying steps of the process that belong to the process. The deterministic behaviour is described using *DeterminismIssue* and a way of tackling it with a use of classes related to the transformation process.

In the third step of the VFramework, which is ”Design verification setting”, the measurement points are specified by designing capture processes and linking them to the elements of the Context Model. The metrics for preservation quality comparison also have their respective classes for expressing the metrics and their value.

In the fourth step of the VFramework, which is ”Capture verification data”, the data is captured from the process by execution of process instances. The information on data location for each of the instances is also covered by the VPlan.

3.3.2 Redeployment environment

The VFramework steps, executed in the redeployment environment, focus on the actual verification of the redeployed process using the information collected in the original environment. At this phase the VPlan is accessed to read the information from it. The Figure 5 depicts which VPlan classes are used at which step of the VFramework. The convention used in the figure is similar to the one from the previous section. The only difference is the direction of the arrows which is opposite, since the information is read from the VPlan.

In the fifth step of the VFramework, which is ”Prepare system for redeployment”, the process is redeployed using information from the process Context Model. The process instances referred to by the VPlan are moved to the system in which they are executed.

In the sixth step of the VFramework, which is ”Capture the redeployment performance data”, the capture process which was used in the original environment is used to capture the information from the redeployed process. Sometimes repetition of the exact capture process is impossible, but it is up to the preservation expert to make a decision how to design a new capture process which is compatible with principles of the original one, which is provided by the VPlan.

In the seventh step of the VFramework, which is ”Compare and asses”, the final assessment of the redeployment is conducted. Information on metrics, their original values and expected values are obtained from the VPlan.

4. VPLAN EVALUATION

In this section we describe the application of the VFramework to an eScience use case. Section 4.1 details the use case. Section 4.2 explains how the VFramework was applied.

4.1 Use Case Description

The safety control of large dams is based on the monitoring of important physical quantities that characterize the structural behaviour (relative and absolute displacements, strains and stresses in the concrete, discharges through the foundations, and so on.). The analysis of data captured by the monitoring systems (sensor networks strategically located at dams) and their comparison with statistical, physical and mathematical models is critical for the safety control assessment. It is known that the variations of hydrostatic pressure and temperature are the main actions that must be considered when analysing the physical quantities generated by the monitoring systems. As a consequence, multiple linear

[7] http://archi.cetis.ac.uk/

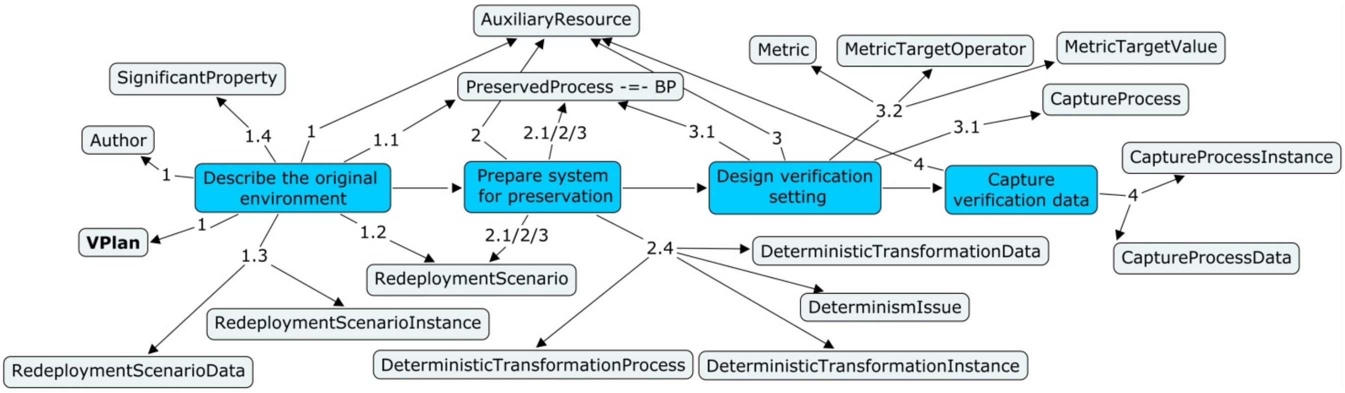


Figure 4: Mapping of the VPlan to the VFramework steps executed in the original environment.

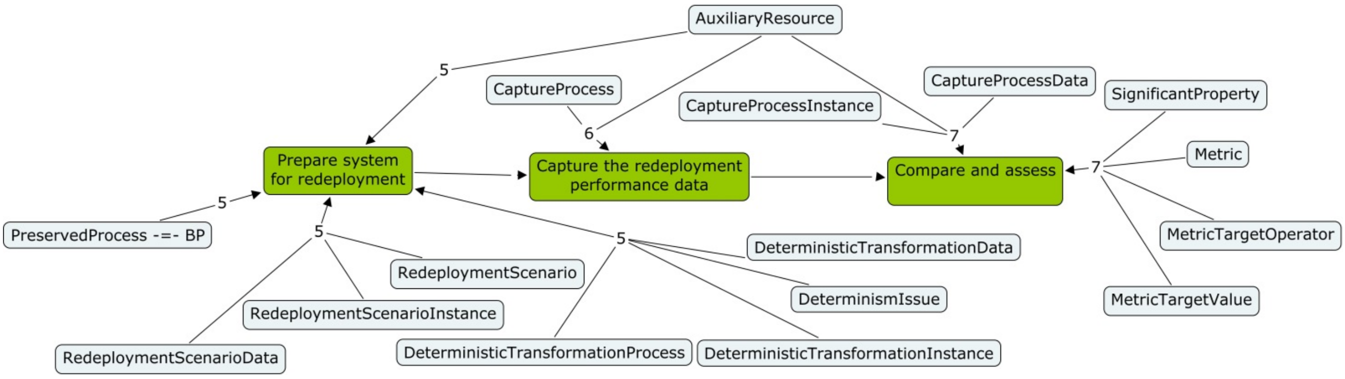


Figure 5: Mapping of the VPlan to the VFramework steps executed in the redeployment environment.

regressions (MLR) are highly suitable and efficient models to determine their relationship with the expected response (physical quantity)[10]. In fact, MLR models are used to model the linear relationship between a dependent variable (predictand or response) and one or more independent variables (predictors).

In large dams, the expected response is approximated by the following effects: (i) elastic effect of the hydrostatic pressures; (ii) elastic effect of temperature, depending on thermal conditions; and (iii) time effect (considered irreversible)[10]. The results of such models are used in structural safety to compare the estimated/predicted behaviour against the real behaviour (represented by the physical quantities captured from the monitoring systems)

Figure 6 details a multiple linear regression process used in dam safety to estimate the physical quantities based on the effects of hydrostatic pressure, temperature and time. For demonstration purposes, this process was isolated from the generic information system (*GestBarragens*). Overall, the process is composed of five steps:

- Extract data: Based on a set of extraction parameters, this process generates the sensor data that will be used in the MLR model (training set with historical values of independent and dependent variables).

- Generate regression: Based on a set of regression parameters (e.g., equation to estimate elastic effect of the hydrostatic pressure), this process generates the regression controls that configure the parameters for the MLR model.

- Execute regression: This process executes the regression parameterized in the regression control, using the training dataset generated in the extract data process. It generates a set of plots and tables to represent the results of the regression execution, including the coefficients (determine the linear relationship between the independent variables and the response, the quality measures (standard deviation, quadratic error, and so on.), residuals (fitting error), and the ANOVA matrix for variance analysis[8].

- Generate aggregation: since a dam has a large number of sensors and a regression is used for each physical quantity associated with each sensor, we might need to run hundreds or thousands of regressions. Thus, the process is able to aggregate all MLR executions into one aggregated report. This step generates the controls that define how this data is aggregated.

- Produce report: This collects all the results produced

[8] The coefficients are used to generate expected responses from the known independent variables. The quality measures, residuals and ANOVA matrix are crucial to determine if a specific MLR model is adequate to estimate and validate a specific physical quantity.

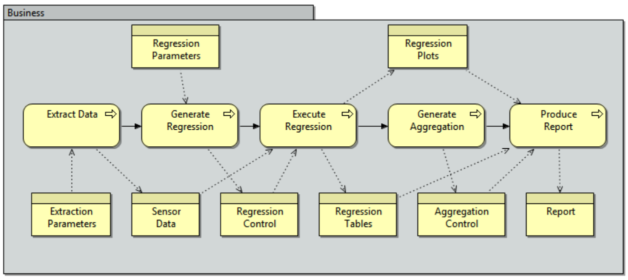


Figure 6: Multiple linear regression process in dam safety.

by the several executions of MLR models and compile them into a single report.

4.2 VFramework Application

As in Section 3.3, we first describe steps taken in the original environment (Section 4.2.1) and then in the redeployment environment (Section 4.2.2).

4.2.1 Original Environment

Following the VFramework, the initial steps have the purpose of collecting all data about the process we want to preserve. This involves initializing a clean ontology file to populate it with the process information. The ontology file will represent the VPlan. In the first step "Describe the original environment" we modelled the process that we want to preserve in ArchiMate using the Archi tool, and imported it to our VPlan. Figure 6 depicts the business layer of the process.

Before import, the process was detailed in terms of the application and technology layer. Note that the final model could also be enriched by the use of context extractors as, for instance, a hardware extractor to further detail the technology layer. It was also defined that the process is preserved with one redeployment scenario in mind. That scenario assumes that the process is fully redeployed to reproduce its original behaviour. One instance of the scenario was stored. Instance data simply consisted of the process application (represented by an executable file at the technological level) and extraction parameters (represented by an "app.config" file) since using the same parameters the application must always produce the same results.

In terms of significant properties that the process needs to maintain we identified and defined the following:

- SP1 - Generate data: the system must be able to generate sensor data for quantitative interpretation.
- SP2 - Export by: the system must generate data for a specific structure, date period and sensor type.
- SP3 - Quantitative interpretation: the system must be able to execute the quantitative interpretation for all the physical quantities of the selected sensor type.
- SP4 - Coefficients: the system must provide the coefficients used in the interpretation, mainly estimate, standard error, t value, Pr(>|t|).
- SP5 - Quality Measures: the system must provide the quality measures of the regression, mainly standard deviation, quadratic error and adjusted quadratic error.
- SP6 - Residuals: the system must provide the residuals of the regression in a table;
- SP7 - ANOVA Matrix: the system must provide the ANOVA matrix of the regression.
- SP8 - Report: the output of the process should be compiled into a single PDF report.

All this information was added to the VPlan. The state of the VPlan after execution of the first step is depicted in Figure 7.

In step two, "prepare system for preservation", the process was analysed in terms of dependencies and determinism. It was concluded that the process is indeed deterministic so there was no need to define a deterministic transformation process. The process has three dependencies on external web-services required to execute the process. We consider that the decision whether to preserve or not the web-services is out of the scope of the VFramework. Ideally stakeholders applying the VFramework should perform a risk analysis to understand whether the web-services are going to be available at redeployment or, if necessary, to preserve them along with the process. In this particular application we did not preserved the web-services and consequently no changes to the VPlan were necessary at this step.

Step three, "design verification setting" is all about assigning metrics to the significant properties and defining how those metrics should be captured. For each metric we defined a text description, a capture process, a target operator and, if applicable, a target value. The combination of the target operator and target value determines the required value of a metric to be considered successful. The absence of the target value indicates that the value of the metric at redeployment should be compared to the value at the original environment. Figure 8 illustrates the definition of a metric using the ontology-editor Protégé[9]. Figure 9 illustrates the capture process entitled "CaptureProcess6" that is defined on Figure 8. All capture processes were defined with the Archi tool, converted to the Context Model and added to the VPlan.

Metrics were associated with significant properties in the following way:

- For SP1, two metrics were defined. Both involve understanding whether "sensor data" generated by the "extract data" step of the process is the same at both the original and redeployment environment. To measure it, one of the metrics involves counting the number of files that were generated and the other consists of counting the number of lines in each file. For the

[9] http://protege.stanford.edu/

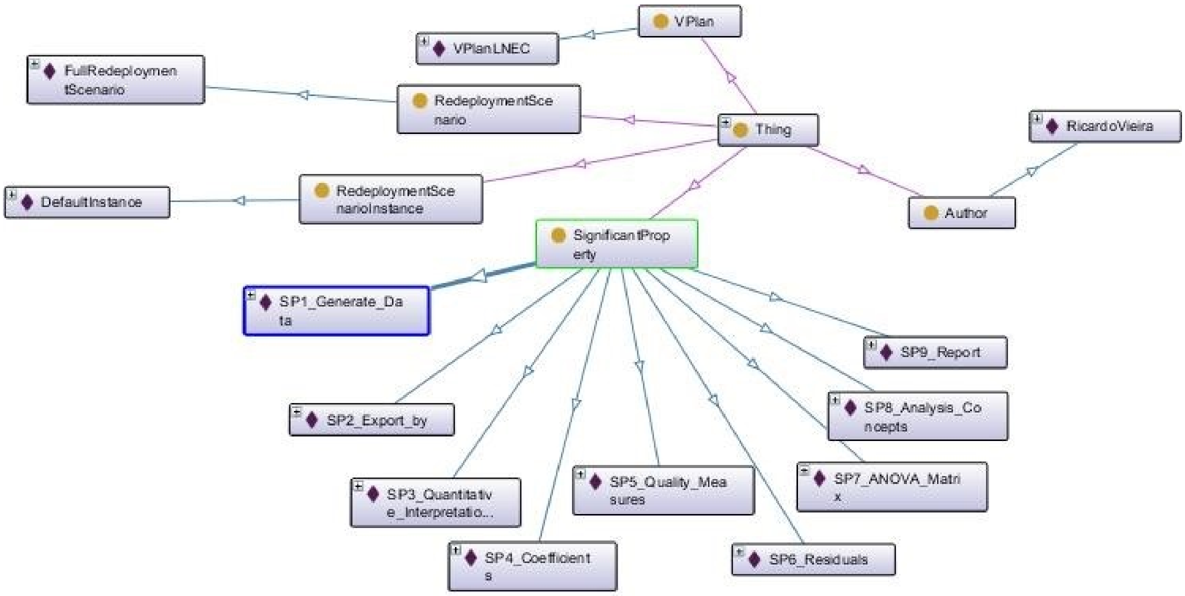
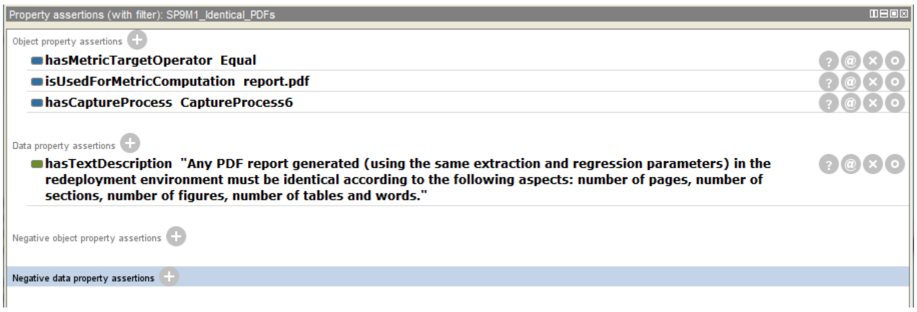


Figure 7: Simplified visualisation of the VPlan after the first step of the VFramework.


Property assertions (with filter): SP9M1_Identical_PDFs

Object property assertions

- **hasMetricTargetOperator Equal**
- **isUsedForMetricComputation report.pdf**
- **hasCaptureProcess CaptureProcess6**

Data property assertions

- **hasTextDescription "Any PDF report generated (using the same extraction and regression parameters) in the redeployment environment must be identical according to the following aspects: number of pages, number of sections, number of figures, number of tables and words."**

Negative object property assertions

Negative data property assertions

Figure 8: Example of a metric modelled in VPlan using Protégé.

same instance of the process, i.e. for each execution of the process using identical ”extraction parameters”, the numbers need to be equal in both environments.

- SP2 had three metrics. Both involve understanding if the generated data conforms to the ”export by” filter. To measure it, we check if the generated data contains data that must not be exported, namely: (1) data from a dam that was not specified; (2) data from a data outside of the selected data period; or (3) data from a sensor not belonging to the selected sensor type.

- SP3 and SP8 had similar metrics. Both properties had one metric and required the execution of the step ”execute regression”. That specific step generates ”regression plots”. SP3 metric involves checking if a plot is generated for each physical quantity present in the sensor data. SP8 metric involves checking if a graphical representation is generated for each analysis concepts (10 concepts in total).

- SP4 to SP7 also have one metric each defined. Again, the capture process involves the execution of the step ”execute regression” but now requires the verification of the generated ”regression tables”. The metrics will verify, respectively, if the ”regression tables” have all coefficients, quality measures, residuals, and ANOVA Matrixes.

- SP9 has one metric to verify if the report generated at the end of the process is equal both in original and redeployment environment. As illustrated in Figure 8 the metric compares the report in terms of number of pages, sections, figures, tables and words.

In the last step at the original environment ”capture verification data” we executed the previous defined capture process and stored the required files. Note that only SP1 and SP9 require comparison between original and redeployment environment so only those capture process were performed at the original enviroment.

4.2.2 *Redeployment Environment*

The fifth step of the VFramework which is ”prepare system for redeployment” involves redeploying the process using the information stored in the VPlan. As in [13], since the preserved process depends on Microsoft .NET Framework 4.0, for redeployment we opted to use a machine running

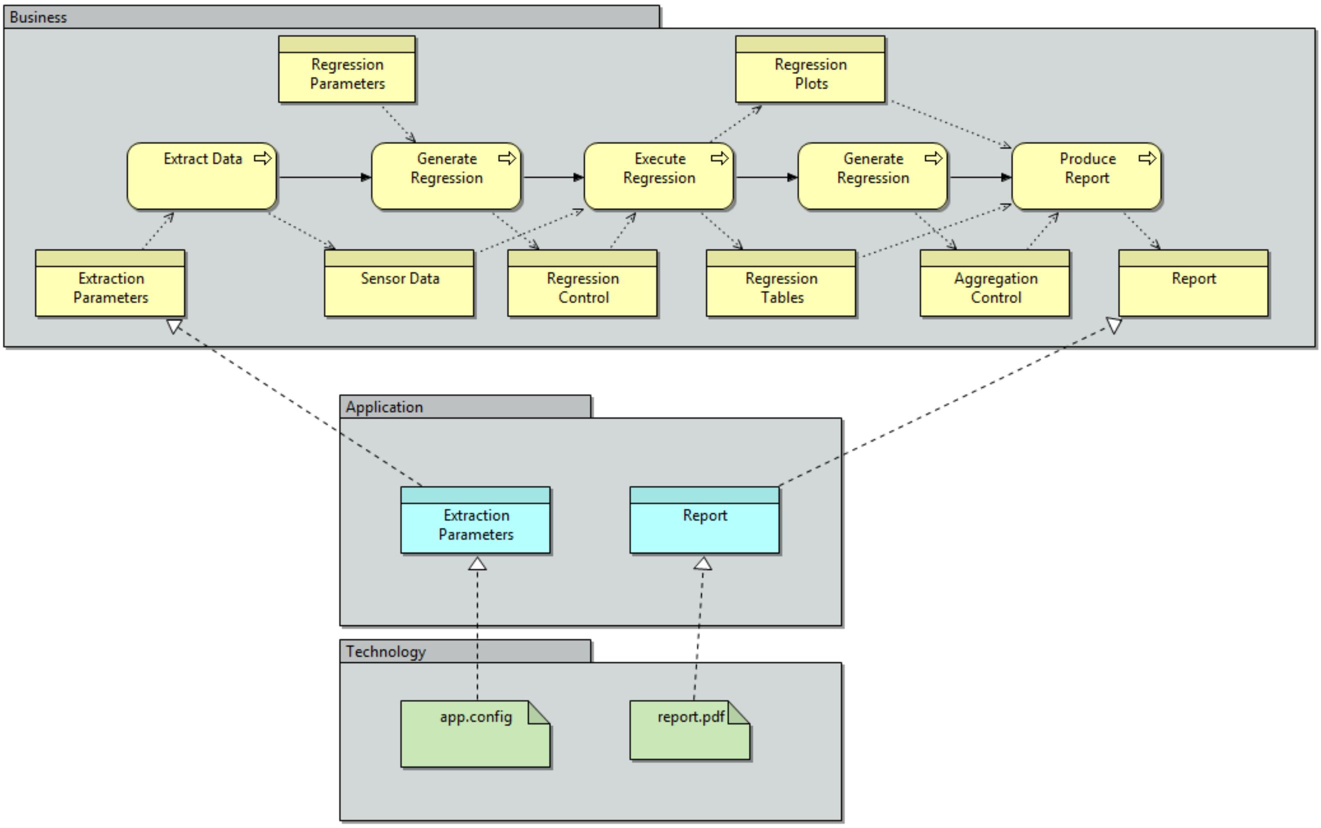


Figure 9: Example of a capture process modelled in Archi.

Ubuntu Linux[10] 12.10 - an open source operating system based on the GNU Linux kernel, which allows us to simulate a slightly different redeployment environment. However, since the .NET platform is exclusively available for Microsoft operating systems, several challenges had to be addressed to re-execute the process in Lunux (for more information refer to [13]).

In the sixth step ”Capture redeployment performance data”, the capture processes defined in the third step ”design verification setting” were executed in the redeployment environment. All processes were executed manually. The result of the execution was a set of files, each associated to a specific metric, that are required for verification of the metric. As an example, the last metric (from SP9) involved executing all the steps of the process and storing the final report for metric assessment in the next step.

Finally, in the last step ”compare and assess” we compared all the results of the capture process to assess if the significant properties were maintained. We consider process to have retained a specific significant property when all of the metrics associated with it are successful verified. To assess a metric we require the target operator and target value (if one exists) from the VPlan in order to understand the type of comparison that needs to be performed and the expected value. All metrics were successfully verified so we concluded that all significant properties from the original environment were maintained at redeployment. Continuing our example, in the metric from SP9 the target operator is ”equal” and there is no target value (as illustrated in Figure 8) meaning that it is necessary to compare data from the original environment (captured in step 5 - ”capture verification data”) with data from the redeployment environment (captured in the previous step). In this specific example we needed to compare two reports, represented as PDF files, in terms of number of pages, sections, figures, tables, and words. Both reports had 25 pages, 5 sections, 80 figures, 33 tables and 1660 words allowing the conclusion that the metric is valid and SP9 was maintained.

5. CONCLUSIONS AND FUTURE WORK

In this paper the VPlan ontology for collection of process verification data was presented. It allows storing information on significant properties, metrics, capture processes and data collected during the verification of preserved and redeployed processes with a use of the VFramework. The VPlan increases the confidence that the evidence needed for the verification of processes is properly organized and stored.

When introducing the VPlan we described its structure (classes and properties) and its integration with the TIMBUS Context Model. Moreover, we provided a mapping of the VPlan concepts to the VFramework in order to demonstrate that the VPlan addresses all of the requirements of the VFramework. Finally, we showed how the VPlan facilitates the ver-

[10] http://www.ubuntu.com/

ification of preserved and redeployed process by applying it to a typical data analysis process from a civil engineering domain.

We are currently working on automation of VPlan creation, so that some of its parts can be automatically generated. This should increase the acceptance within the scientific community. We are also developing a set of SPARQL queries which not only validate the VPlan, but also facilitate retrieval of the information stored in the VPlan. Future work will also focus on further testing on different use cases.

ACKNOWLEDGMENTS

This research was co-funded by COMET K1, FFG - Austrian Research Promotion Agency and by the European Commission under the IST Programme of the 7th FP for RTD - Project ICT 269940/TIMBUS.

6. REFERENCES

[1] R. S. Aguilar-Savén. Business process modelling: Review and framework. *International Journal of Production Economics*, 90(2):129 – 149, 2004. Production Planning and Control.

[2] G. Antunes, M. Bakhshandeh, R. Mayer, J. Borbinha, and A. Caetano. Using ontologies for enterprise architecture analysis. In *Proceedings of the 8th Trends in Enterprise Architecture Research Workshop (TEAR 2013), in conjunction with the 17th IEEE International EDOC Conference (EDOC 2013)*, Vancouver, British Columbia, Canada, September 9-13 2013.

[3] K. Belhajjame, O. Corcho, D. Garijo, J. Zhao, P. Missier, D. Newman, R. Palma, S. Bechhofer, E. Garcia Cuesta, J. M. Gomez-Perez, S. Soiland-Reyes, L. Verdes-Montenegro, D. De Roure, and C. Goble. Workflow-centric research objects: First class citizens in scholarly discourse. In *Proceedings of Workshop on the Semantic Publishing*, 2012.

[4] C. Collberg, T. Proebsting, G. Moraila, A. Shankaran, Z. Shi, and A. Warren. Measuring Reproducibility in Computer Systems Research. Technical report, 2013.

[5] M. Guttenbrunner and A. Rauber. Evaluating an emulation environment: Automation and significant key characteristics. In *Proceedings of the 9th International Conference on Digital Preservation (iPres 2012)*, pages 201–208, Toronto, Canada, October 1-5 2012.

[6] M. Guttenbrunner and A. Rauber. A measurement framework for evaluating emulators for digital preservation. *ACM Transactions on Information Systems (TOIS)*, 30(2), 3 2012.

[7] V. Haren and V. H. Publishing. *ArchiMate 2. 0 Specification*. The Open Group. Van Haren Publishing, 2012.

[8] K. Hettne, S. Soiland-Reyes, G. Klyne, K. Belhajjame, M. Gamble, S. Bechhofer, M. Roos, and O. Corcho. Workflow forever: Semantic web semantic models and tools for preserving and digitally publishing computational experiments. In *Proceedings of the 4th International Workshop on Semantic Web Applications and Tools for the Life Sciences*, SWAT4LS '11, pages 36–37, New York, NY, USA, 2012. ACM.

[9] S. Jones. A report on the range of policies required for and related to digital curation. Technical Report 1, Mar. 2009.

[10] J. Mata. Interpretation of concrete dam behaviour with artificial neural networks and multiple linear regression models. *Engineering Structures*, 33(3):903–911, 2011.

[11] R. Mayer, A. Rauber, M. A. Neumann, J. Thomson, and G. Antunes. Preserving scientific processes from design to publication. In P. Zaphiris, G. Buchanan, E. Rasmussen, and F. Loizides, editors, *Proceedings of the 16th International Conference on Theory and Practice of Digital Libraries (TPDL 2012)*, volume 7489 of *Lecture Notes in Computer Science*, pages 113–124, Cyprus, September 23–29 2012. Springer.

[12] B. D. Mccullough. Got Replicability? The Journal of Money, Credit, and Banking Archive. *Econ Journal Watch*, 4(3):326–337, Sept. 2007.

[13] T. Miksa, S. Proell, R. Mayer, S. Strodl, R. Vieira, J. Barateiro, and A. Rauber. Framework for verification of preserved and redeployed processes. In *Proceedings of the 10th International Conference on Preservation of Digital Objects (IPRES2013)*, Lisbon, Portugal, September 2–6 2013.

[14] T. Miksa and A. Rauber. Increasing preservability of research by process management plans. In *Proceedings of the 1st International Workshop on Digital Preservation of Research Methods and Artefacts*, DPRMA '13, pages 20–20, New York, NY, USA, 2013. ACM.

[15] P. Missier, S. S. Sahoo, J. Zhao, C. A. Goble, and A. P. Sheth. Janus: From Workflows to Semantic Provenance and Linked Open Data. In *Proceedings of the International Provenance and Annotation Workshop (IPAW2010)*, pages 129–141, Troy, New York, USA, June 15–16 2010.

[16] O. M. G. (OMG). Business process model and notation (bpmn) version 2.0. Technical report, jan 2011.

[17] S. Strodl, D. Draws, G. Antunes, and A. Rauber. Business process preservation, how to capture, document & evaluate. In *Proceedings of the 9th International Conference on Preservation of Digital Objects (IPRES2012)*, Toronto, Canada, October 2012.

[18] S. Strodl, R. Mayer, D. Draws, A. Rauber, and G. Antunes. Digital preservation of a process and its application to e-science experiments. In *Proceedings of the 10th International Conference on Preservation of Digital Objects (IPRES 2013)*, 9 2013.

[19] J. S. Susanne Glissman. A comparative review of business architecture. Technical report, IBM Research Division, August 24 2009.

[20] The Open Group. *Archimate 2.0: A Pocket Guide*. TOGAF series. Van Haren Publishing, 2012.

[21] TIMBUS Consortium. D4.6: Use Case Specific DP & Holistic Escrow. Technical report, 2013.

Occam’s Razor and File Format Endangerment Factors

Heather Ryan
University of Denver
Library & Information Science Program
1999 E. Evans Avenue
Denver, CO 80208
heather.m.ryan@du.edu

ABSTRACT

Much digital preservation research has been built on the assumption that file format obsolescence poses a great risk to the continued access of digital content. In efforts to address this, a number of researchers created lists of factors that could be used to assess risks associated with digital file formats. This research examines these assumptions about file format obsolescence and file format evaluation factors with the aim of creating a simplified file format endangerment index.

This study examines file format risk under a new lens of file format endangerment. Using the Delphi method in two separate studies, this exploratory research collected expert opinion on relevance of a list of factors as causal indicators of file format endangerment.

The findings show that only three of the dozens of file format evaluation factors discussed in the literature exceeded an emergent threshold level as causes of file format endangerment: *rendering software available*, *specifications available*, and *community/3rd party support*. These factors are ideal candidates for use in a file format endangerment index.

General Terms

infrastructure, communities, strategic environment, preservation strategies and workflows

Keywords

endangerment, file formats, formative measurement model, obsolescence

1. INTRODUCTION

Occam’s Razor is “a scientific and philosophic rule that entities should not be multiplied unnecessarily which is interpreted as requiring that the simplest of competing theories be preferred to the more complex or that explanations of unknown phenomena be sought first in terms of known quantities” [1]. The principle of Occam’s Razor can be broadly translated into the notion that it is better to solve problems using the simplest solution.

This study, and its findings, calls into question the notion that assessing file format risk should involve complicated models with dozens of calculated and weighted evaluation factors. A conversation started by Johan van der Knijff [2][3] on the Open Planets Foundation website points out that many of the factors included in these models are theoretical, untested, and sometimes not testable. I agree.

Through the research I present here, I (and my study participants) have taken Occam’s Razor to the dozens of file format evaluation factors found in the literature. I introduce a formative measurement model, i.e., an index, as the framework to guide a more exact method of selecting a simple set of file format endangerment factors.

Within the context of this research, I also propose a shift in language usage from *obsolescence* to *endangerment*. *File format obsolescence* is a phrase commonly used to describe the phenomenon that occurs when information stored in a particular file format is no longer accessible using current technology. Although it has often been the focus of research and discussion

While the term *file format obsolescence* is still useful to describe a state in which a file format is no longer in use, I will use the term *file format endangerment* to describe the possibility that information stored in a particular file format will not be interpretable or renderable using standard methods within a certain timeframe. This term will be used in a way that is similar to its application to animal species. According to Merriam-Webster, *endanger* means, “to bring into danger or peril,” where an endangered species is “a species threatened with extinction,” or more broadly, “anyone or anything whose continued existence is threatened” [1]. A file format is not threatened with extinction or a discontinued existence; rather the threat is to the ability to access information from a file that is encoded in that format.

Using the phrase *file format endangerment* provides a new perspective for studying the nature of these risks. By studying a file format’s ability to be rendered as being similar to animal species endangerment, potentially useful parallels may be created that can lend new insight into the problem. Animal species have been studied for hundreds of years, and the methods used to document and assess the factors that contribute to their thriving or extinction can be applied to the viability or inaccessibility of the different “species” of file formats. From this we can learn which factors most heavily contribute to the risk of file format endangerment, and we can use this knowledge to identify this risk and take action to ameliorate it. Finally, the term “endangerment” embodies a sense of hope and urgency that hopefully incites action; much more so than the term obsolescence, which emits a sense of loss that is irreparable.

2. LITERATURE REVIEW

I explored the literature to identify and review past and present initiatives in file format risk evaluation, lists of file format evaluation factors, and measurement models that could be used to guide file format evaluation.

2.1 Initiatives in File Format Risk Evaluation

Several projects have approached the process of file format risk assessment and notification. These are the Automated Obsolescence Notification System (AONS), AONS II, parts of the Archive Ingest and Handling Test (AIHT), Plato, Scout, and research conducted at the Austrian Institute of Technology.

AONS[1] was a project of the National Library of Australia (NLA) and the Australian Partnership for Sustainable Repositories (APSR) and built upon work of the Preservation Architecture for New Media and Interactive Collections (PANIC) project, discussed later. In 2006, AONS was developed to create a file format obsolescence alert system, specifically for the DSpace digital repository platform. The alert system was to be built on an architecture that used DROID for file format identification, and PRONOM and Library of Congress Directory of Formats to provide obsolescence risk evaluation. If file formats found in the repository are identified to be at risk, the system generates a risk report and sends the report to the repository manager [4].

In 2007, work on AONS II began in order to refine the AONS services. Notably, the AONS II report stated, "an initial business driver for the project was a perceived need for a tool which could automate much of the assessment process, using standardized metrics that would support machine-formulation of recommendations on risk levels" [5]. Unfortunately, the project relied heavily on risk reporting capabilities of PRONOM, which have yet to come to fruition.

The Archival Ingest and Handling Test (AIHT) project[2] (2004-2005) was funded by the Library of Congress to "assess the digital preservation infrastructures of four small, real-world digital archives" [6]. The four partners were Johns Hopkins University, Sheridan Library; Harvard University Library; Old Dominion University Department of Computer Science; and Stanford University, Libraries and Academic Information Resources (Library of Congress, n.d.). As part of the AIHT, the Stanford University participants developed a file format risk-assessment system. They based their system on JHOVE for file format identification and representation information and the Arms and Fleischhauer [7] list of preferred file formats, from which they created a matrix for risk-assessment. From this they developed what they called the Empirical Walker Process, intended to be a fully automated metadata and risk-assessment generator that flags materials that may be in danger of becoming obsolete [6].

After developing this prototype system, Anderson, Frost, Hoebelheinrich, and Johnson evaluated the resources required to automate and maintain a preservation assessment of the Empirical Walker Process, such as maintaining the infrastructure to support the process. While they have yet to fully develop this process, they suggested that the cost to manage such a system was too much for one institution to bear and suggested, "perhaps a federated approach to some of this activity, as a service to a community of repositories and their users, would be most economical" [6].

Plato[3] (2005-present) was developed as part of the Planets preservation-planning project. Plato addresses many aspects of preservation planning [8]. Among them is assessing file format criteria that could indicate risk. They propose to evaluate file formats based on the criteria: browser support, standardization, ubiquity, stability, licensing, compression, format documentation, tool support, comparative file size, complexity, disclosure, master can be used as access copy, Optical Character Recognition (OCR applicable, and adoption. Becker and Rauber cite several obstacles toward realizing the goal of automating the process of measuring and evaluating formats based on these criteria: 1. only roughly 20% of the criteria can be automatically measured, 2. external sources of data or not complete and, 3. there is a lack of standardized benchmarks that can be used in comparative analysis.

Scout[4] is a semi-automatic preservation watch system being developed within the Scalable Preservation Environments (SCAPE) project (2011-present), "an EU-funded project which is directed towards long term digital preservation of large-scale and heterogeneous collections of digital-objects" [9]. Scout was designed to collect information from various sources that can be used to detect risks to digital content. It collects information from various registries like PRONOM as well as through natural language extraction from the World Wide Web [10][11]. This tool is still under development and has undergone only basic, proof-of-concept testing.

Another, similar approach toward file format risk analysis is being developed by Roman Graf and Sergiu Gordea (2011-present) [12][13], both of the Austrian Institute of Technology. They are also developing a system that collects data from various sources to analyze file formats for what they call, "preservation friendliness." They designed their system to collect data from PRONOM, DBPedia, and Freebase on twenty-one identified risk factors. They collected and analyzed data for these factors for a set of thirteen representative file formats to produce a total risk percentage value for each file format.

A few groups have developed digital preservation systems that incorporate file format risk analysis into workflows. These are the Preservation Services Architecture for New media and Interactive Collections (PANIC), Ex Libris' Rosetta, Tessella's Safety Deposit Box, and the National Library of the Netherland's (KB) *e*-Depot.

PANIC[5] (2004-2006) is a "semi-automated digital preservation system based on semantic web services" [14]. The project, funded by the Cooperative Research Centre for Enterprise Distributed Systems Technology (DSTC) and the Australian Federal Government's CRC Programme, facilitated the building of a prototype system to assess a digital object's obsolescence risk and subsequently invoke migration or emulation tools to counteract the risk. The system architecture contains invocation, notification, discover, and provider components. The invocation component was designed to detect obsolescence using information retrieved from the built-in software version registry via a notification agent. This registry contains information about software that is used to render the objects in the collection. Once notified of risk, the discovery component is set into action to locate appropriate preservation services using the OWL-S ontology that is used for describing and discovering web services. The provider component then sends the at-risk files to the located service that then performs the requested service [15]. There has been no development of PANIC beyond the prototype phase.

[1] apsr.anu.edu.au/aons

[2] www.digitalpreservation.gov/partners/aiht.html

[3] www.ifs.tuwien.ac.at/dp/plato/intro

[4] openplanets.github.io/scout

[5] www.itee.uq.edu.au/eresearch/projects/panic

Rosetta[6] (2009-present) is a digital preservation system produced by the Ex Libris Group [16][17]. The system has a deposit module, a working area, a permanent repository module, an operational repository, a preservation planning module, an administration module, and an access module. According to the software description, the preservation-planning module provides risk analysis of file formats, but there is no indication as to how this is accomplished. I contacted a representative of Ex Libris who stated that due to the proprietary nature of their product, they could not share information beyond what is available online.

Safety Deposit Box (SDB)[7] (2011-present) is part of the Preservica digital preservation suite developed by Tessella [18]. Key features of SDB are ingest, data management, storage, access, preservation planning and action, and administration. The preservation planning and action feature uses file characterization tools to assess file format risk, though there is no clear source of internal or external file format risk information and no clear evidence that this function is operational. As of this writing, the file format evaluation component of SDB is still not production ready, though, "Tessella are moving to a 'linked data' registry in the next release. The plan is to revisit the ability to define a format risk assessment in a future release once the linked data version is stable" (Evans, M., personal communication, January 24, 2014).

e-Depot[8] (2004-present) is a system built for the National Library of the Netherlands using the IBM system, Digital Information Archiving System (DIAS) [19]. DIAS was extended to include a Preservation Subsystem that included a functionality called the Preservation Manager that stores technical metadata that specifies the software and hardware necessary to render the file formats stored in *e*-Depot. This functionality was designed to meet three objectives: "1) Identify[ing] the electronic publications in danger of becoming inaccessible due to technology changes, 2) Planning the activities associated with preservation, i.e. implementing migration and/or emulation strategies, and 3) Specifying the software and hardware environments required to render an electronic publication" [19]. At the time of this writing, the KB web page on *e*Depot states that, "Preservation functionality will be enhanced in future DIAS versions to generate signals when stored assets must be converted or migrated to ensure their availability" [20]. Attempts to communicate with representatives from the KB to learn more yielded no results.

Digital preservation researchers and developers have put a great deal of work into creating tools and systems designed to manage and preserve digitally encoded information. A close examination of the existing tools, however, reveals a gap in a critical area of need: none of these tools and systems operationally addresses the issue of file format risk monitoring, though some developers claim their systems do or will do in the future. Many of the tools and systems discussed here claim that their file format risk analysis components will come from PRONOM, but while PRONOM has a place for it in its data model, it does not currently contain information on file format risk information. In fact, none of the tools or systems listed here has proven functionality in file format risk analysis. This shows that though the digital preservation community indicates that it is important to monitor file format risk, they have yet to find a viable way to do this.

It is not entirely clear what is preventing further progress in this area, but one obvious needed improvement is to flesh out the existing collections of file format data. Because so many of the tools and systems discussed here rely on sparse and non-existent data in the file format registries, their full functionality is hindered. Beyond this, a more clear understanding of which factors should be measured to provide proposed risk ratings will allow the community to focus its data collection efforts on the most useful and beneficial information. Before factor can be chosen and before data can be collected, it is imperative to have a clear understanding of which model to use to shape the development of a trustworthy file format endangerment measure.

2.2 Formative Indicators and Index Construction

Conservation biology and file format endangerment both involve the collection and analysis of data for pre-defined factors to detect potential dangers. The pre-defined factors represent indicators of the phenomenon being measured, i.e., species endangerment, epidemics, or file format endangerment; and are commonly called *formative indicators*.

Formative indicators, used in index construction, have an opposite relationship than do "effect" or "reflective indicators," which are commonly used in scale development. The opposite causal directions of reflective and formative measurement models are illustrated in Figure 1, where η is the construct or phenomenon being measured, and x_1, x_2, and x_3 are the reflective and formative indicators. In panel 1, λ represents the relationship that the construct has on the reflective indicators, x_1, x_2, and x_3. The symbol ε represents the error. In panel 2, ζ is a disturbance term that represents remaining relationships of the construct that are not represented by the formative indicators and that cannot be measured. The symbol γ represents the relationship that the formative indicators, x_1, x_2, and x_3 have on the construct and the *r* variables and their incumbent arrows represent their interdependency toward defining, creating, and indicating causes of the construct.

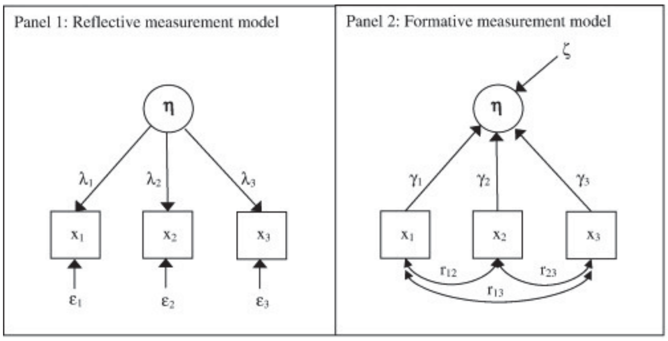

Panel 1: Reflective measurement model
η
λ1
λ2
λ3
x1
x2
x3
ε1
ε2
ε3
Panel 2: Formative measurement model
ζ
η
γ1
γ2
γ3
x1
x2
x3
r12
r23
r13

Figure 1. Causal direction in reflective and formative measurement models [21].

As an example of a formative measure, the construct or the phenomenon that I intend to measure is file format endangerment. The formative indicators are the factors that are determined to indicate causes of file format endangerment. In a reflective measure, the effects, i.e. the reflective indicators of the phenomenon, are measured, such as in personality measures where the personality is the construct and the personality traits are measured as an effect of the personality. According to Bollen,

[6] www.exlibrisgroup.com/category/RosettaOverview

[7] tessella.com/products/tessella-sdb

[8] www.kb.nl/en/expertise/e-depot-and-digital-preservation

"most researchers in the social sciences assume that indicators are effect indicators," where, "cause indicators are neglected despite their appropriateness in many instances" [22].

It is often not clear or obvious which of the two measurement models is most appropriate. Bollen [22] suggests that one method of determining which model is more appropriate is to perform a "temporal priority" mental experiment, or simply put, think about which happens first: the indicator or the construct. In the case of file format endangerment, my intention was to create a predictive model using factors that precede endangerment. Consequently, such a model demonstrates the temporal priority of factors that are exhibited before the phenomenon of file format endangerment. Phenomenon prediction requires data collection for *a priori* factors, or observable factors that occur before the measured phenomenon; therefore, a formative measurement model best suits the purposes of evaluating the possibility that information encoded in a particular file format will become inaccessible within a certain timeframe.

Once a researcher has determined that the indicators in question have a formative relationship with the construct, they can begin to design the measurement model, or index. Diamantopoulos and Winklehofer [23] describe the four steps for constructing an index:

1. Content Specification - defining the "domain of content the index is intended to capture"
2. Indicator Specification - choosing the indicators to be added to and tested for the index.
3. Indicator Collinearity - checking that there is not excessive collinearity between the indicators.
4. External Validity - determining that the index measures what it claims to measure and "assessing the suitability of the indicators"

Diamantopoulos and Winklehofer suggest that the definition of the domain be broad enough to encompass all of the causal indicators. Though they provide no formal recommendation for specifying which indicators to include in an index, they reported that they selected indicators for their export market sales forecasting index through "an extensive review of the forecasting literature as well as exploratory interviews with export managers" [23].

In respect to indicator collinearity, formative indictors in indexes should have a direct effect on the phenomenon being measured and have little to no intercorrelation, meaning the indicators in a formative measure should have little to no direct effect on each other. While indicators in a formative measure may have some interaction with each other, it is best if they do not have strong correlations with one another [24].

Finally, determining external validity involves testing the index to determine if it measures the specified construct. Diamantopoulos and Winklehofer suggest, "One possibility is to use as an external criterion a global item that summarizes the essence of the construct that the index purports to measure" [23].

The research presented here addresses the first two of the above steps. For the first step, I specify the content of the file format endangerment index as being all factors that indicate a cause, either through their presence or absence, information encoded in particular file formats to become inaccessible over a specified timeframe. Similar to Diamantopoulos and Winklehofer, I addressed indicator specification through an extensive literature review, supplemented by the factor-rating Delphi exercise described below. I intend to address steps three and four in future research.

2.3 File Format Evaluation Factors in the Literature

Effective analysis of file format endangerment requires a well-constructed and validated index to guide data collection. The key to creating a valid index is choosing the right factors that have a formative relationship with the measured phenomenon. Previously, researchers from various institutions created several different lists of file format evaluation criteria. Some of these lists of criteria were designed to evaluate aspects of file formats that can contribute to or alleviate risks associated with file formats. While none of these lists were created with the intention of creating a file format endangerment index, the approaches used are similar enough to provide a useful starting point for the index development process.

At the beginning of this research process, I identified twelve sets of file format evaluation criteria from the literature listed in Table 1. Within these lists, I identified 138 individual factors. The lists have varying numbers of factors. Some had as few as five factors, and one had as many as 22.

Table 1. Sources of File Format Evaluation Factors

Project/Program/Institution	Year
Risk Management for Digital Information Project; Council on Library and Information Resources [25]	2000
Math*Diss* International Project and EMANI project; Niedersächsische Staats- und Universitätsbibliothek, Götingen [26]	2003
Groupe Pérennisation des Informations Numériques (PIN) [27]	2004
Internetbevaringsprojektet (the Internet Preservation Project); Statsbiblioteket (The State Library), Det Kongelige Bibliotek (Royal Library, Denmark) [28] [29]	2004
INvestigation of Formats based on Risk Management (INFORM) [30]	2004
Automated Preservation Assessment of Heterogeneous Digital Collections (AIHT) [31] [32]	2005
The National Archives (TNA-UK) [33] [34]	2005
Service Oriented Architecture (SOA); University of Minho, Portugal [35] [36]	2006 2007
International Research on Permanent Authentic Records in Electronic Systems 2 (InterPARES) [37]	2007
National Centre for Radio Astrophysics [38]	2007
Koninklijke Bibliotheek (KB) (The Royal Library, Netherlands) [39]	2008
Preservation and Long-term Access through Networked Services (PLANETS) [40]	2008

3. RESEARCH METHOD

One of the primary objectives of this research was to clarify which of the many factors discussed in the literature are the most relevant formative indicators to include in a file format endangerment index. The research described here took a three-pronged approach to addressing these issues: two separate Delphi studies and one information gathering and rating exercise designed to test a unification of the two Delphi studies.

The Delphi method was the most effective method to determine which are the most relevant factors that indicate a cause of file format endangerment. When little data exists on a topic, such as with file format endangerment, Delphi is known to be an effective method of “producing trustworthy personal probabilities regarding hypotheses” in experts’ knowledge area [41]. Dalkey [42] explained that characteristics of a Delphi procedure are anonymity, iteration with controlled feedback, and statistical group response. These procedures were designed to reduce “the influence of certain psychological factors, such as specious persuasion, the unwillingness to abandon publicly expressed opinions, and the bandwagon effect of majority opinion.” Gordon and Helmer suggested that inviting participants to review other panel members’ reasoning will promote a thoughtful consideration of ideas and will lead to a more accurate representation of the truth. [43]

After performing Bollen’s [22] temporal priority mental experiment, described in Section 2.2, I determined that the factors I was examining for file format endangerment occurred before the phenomenon of file format endangerment. This pre-phenomenal occurrence indicates that the factors should be considered as potential causal indicators of file format endangerment, and thus appropriate for use in an index.

3.1 Selecting File Format Endangerment Factors for Review

My review of existing literature revealed many discussions of the importance of assessing a file format’s stability for long-term preservation. Several of these discussions include proposed measures for assessing file formats for preservation purposes, as discussed in the literature review. I used these lists as the starting point for what eventually became the list of file format endangerment factors rated in the Factor Rating Questionnaire.

I used a semi-structured method to compile a draft list of factors. I copied each of the evaluation criteria into a document with citations to the original reports for reference. I then compiled all of the factors into one list, removing exact duplicates as I went. This process resulted in a list of nearly fifty factors.

I then started a new list of factors, grouping similar factors together by reviewing provided descriptions. For example, I grouped *widely accepted*, *widespread use*, *popularity*, *market share*, and *adoption* under the factor *ubiquity*. I evaluated each group of similarly themed factors and selected a name for the group that best described them. I made no value judgments as to the factors’ viability as formative indicators of file format endangerment. This process resulted in a list of twenty factors. I then wrote definitions for each of the remaining factors.

I provided a list of all of the factors that were presented in the literature to a knowledgeable colleague who independently performed the same task. There were a number of differences in the way this person grouped and named the factors. We met and discussed each of our factor groupings and reached an agreement on the final synthesis of factor lists. The following are the resulting factors and their definitions:

Backward/Forward Compatibility - whether or not newer versions of the rendering software can render files from older versions, or whether or not older versions of rendering software can render files from newer versions.

Community/3rd Party Support - the degree to which communities and/or parties beyond the original software producers support the file format.

Complexity - relates to how much effort has to be put into rendering and understanding the contents of a particular file format.

Compression - whether or not, and the degree to which a file format supports compression.

Cost - The cost to maintain access to information encoded in a particular file format, e.g. to migrate files, to maintain the rendering software, or to run an emulation environment.

Developer/Corporate Support - whether or not the entity that created the original software that produces output in the file format continues to support it.

Ease of Identification - the ease with which the file format can be identified.

Ease of Validation - the ease with which the file format can be validated, where validation is the process by which a file is checked for the degree to which it conforms to the format’s specifications.

Error-tolerance - the degree to which this format is able to sustain bit corruption before it becomes unrenderable.

Expertise Available - the degree to which technological expertise is available to maintain the existence of software that can render files saved in this format.

Legal Restrictions - the degree to which this file format is or can be restricted by legal strictures such as licensing, copy and intellectual property rights.

Lifetime - the length of time the file format has existed.

Metadata Support - whether or not the file format allows for the inclusion of metadata.

Rendering Software Available - whether or not any type of software is available that can render the information stored in this file format.

Revision Rate - the rate at which new versions of this file format’s originating software are released.

Specifications Available - whether or not documentation is freely available that can be used to create or adapt software that can render information stored in this file format.

Standardization - whether or not this file format is recognized as a standard for use and/or preservation by a reputable standards body.

Storage Space - the average amount storage space a file saved in this format requires when saved.

Technical Dependencies - the degree to which this file format depends on specific software, operating systems, and hardware in order for its contents to be successfully accessed or rendered.

Technical Protection Mechanism - whether or not this file format allows for or is encumbered by technical protection mechanisms such as Digital Restrictions Management (DRM).

Ubiquity - the degree to which use of this file format is widespread and in common use.

3.2 Research Design

This research involved the use of four questionnaires; administered online using Qualtrics survey software:

1. A questionnaire designed to collect information about the quantity and quality of experience that recruited Delphi participants had working with file formats in a digital preservation context. I used the information collected from

this questionnaire to determine the expertise level of participants and to assign them to one of the two Delphi groups.

2. A questionnaire designed to collect information on participant opinions of file format endangerment level ratings of 50 test file formats. I administered this questionnaire in a Delphi process in which participants answer the questionnaire over multiple rounds and review anonymous responses of their fellow participants between rounds.
3. A questionnaire designed to collect information on participant opinions of the relevance of factors as a cause of file format endangerment.
4. A questionnaire designed for one special rater participant to collect and report on information about factors for a list of file formats, to collect endangerment level ratings for the list of file formats, and to collect relevancy ratings for the list of factors considered as causes of file format endangerment. I designed this exercise to provide an additional source of data collection for both understanding the current perceived level of file format endangerment and for understanding which factors are direct causes of file format endangerment.

The results presented here are focused primarily on the third and fourth questionnaire. In the third questionnaire, I presented participants with the list of file format evaluation factors compiled from the dozen file format evaluation lists found in the literature.

In this questionnaire, I asked participants to rate each factor on an ordinal scale that indicates degrees of relevancy of the factor as a cause of file format endangerment:

- Not relevant at all
- Somewhat relevant
- Very relevant

I also asked participants to provide a brief narrative to explain their ratings for each of the factor options. Additionally, I asked participants to suggest factors that they believed to be a cause of file format endangerment that were not included in the original list, and their rational for suggesting the factors.

After participants completed their questionnaires, I created a document with participants' anonymized ratings and explanatory narratives for each questionnaire. I shared this document with participants and asked them to review each other's answers and narratives, and to thoughtfully reconsider their original answers. I then asked them to answer a fresh version of the questionnaire in a second round.

Some participants suggested additional index factors during the first round of the Factor Rating Questionnaire. I reviewed the 16 suggested factors, and from them, selected six new factors that had not in some way been addressed by the original list of 21 factors. For example, one participant suggested, "Existence of a community around the format," however, this factor was already addressed under the factor, *community/3rd party suppor*t.

Additionally, I evaluated the justification narratives in the first round of the Format Rating Questionnaire for the emergence of additional factors that should be included in the Factor Rating Questionnaire. Based on this evaluation, I added the factor, *value* to the second round of the Factor Rating Questionnaire. I added a total of seven new factors to the second round of the Factor Rating Questionnaire and asked participants to rate them on the same scale as the original twenty-one factors. The following are the seven new factors that I added to the original 21 factors to be rated in Questionnaire 2, Round 2:

Value - the degree to which information encoded in this format is valued.

Geographic Spread - the way in which a file format is spread across the world; whether spread thinly across the globe or condensed heavily in a particular area.

Domain Specificity - the degree to which the format is used only within specific domains.

Viruses - the degree to which the format is susceptible to containing or being damaged by viruses.

Availability Online - the degree to which the format is available on the Web.

Institutional Policies - the degree to which a file format is affected by institutional polices, such as whether or not an institutional policy states that content encoded in this format will be collected and preserved.

Specification Quality - (sub-factor of "Specifications Available") the understandability and usefulness of the format's available specifications in maintaining access to content encoded in that format.

I asked participants to answer the Factor Rating Questionnaire for a third time with only the seven new factors introduced in the second round. This gave participants an opportunity to rate the new factors a second time. As with previous rounds, I collected the anonymized responses into a document and asked participants to review the document as they re-rated the factors. After the second round of rating for each factor, I determined that participant ratings had not changed substantially enough to continue to additional rounds.

The fourth questionnaire was administered to one trained, special reviewer. In this questionnaire, the reviewer was presented with each of the file formats that were used in the Format Rating Questionnaire. For each file format, I asked the reviewer to:

1. Review a guide on possible data collection sources that I created based on data I collected from the file format rating Delphi questionnaire.
2. Collect and share information from online sources, other recommended sources, or from personal knowledge for each of the factors selected during data analysis of the Factor Rating Questionnaire.

After considering the data collected in step 2, I then asked the reviewer to rate each file format on the file format endangerment level scale used in the Format Rating Questionnaire:

- Information stored in this file format is already inaccessible.
- Information stored in this file format will be inaccessible in 1-5 years.
- Information stored in this file format will be inaccessible in 6-10 years.
- Information stored in this file format will be inaccessible in 11-20 years.
- Information stored in this file format will be inaccessible in 20 years or more.
- I am not familiar enough with this file format to rate it.

After the reviewer collected factor information for each of the forty-three file formats, I asked him to rate each of the factors using the same scale for relevancy as a cause of file format endangerment that was used in the Factor Rating Questionnaire:

- Not relevant at all

- Somewhat relevant
- Very relevant

Because the special rater had just gone through the exercise of searching for information on each factor and applying this directly to rating the file formats, his ratings were strongly based in the reality of putting the factors to use in a real-world scenario. This activity provided me with additional data that I used to compare with other factor-related data that I collected from the file format rating and factor rating Delphi questionnaires.

I conducted a semi-structured e-mail interview in which I elicited feedback on the process the special reviewer used to collect information for each factor, how useful he found each factor to be in assessing file format endangerment, and any other thoughts and opinions he had about the process.

3.3 Participants

I selected participants for the two Delphi questionnaires from a group of individuals I identified as having expertise on file formats. Luo and Wildemuth recommended that experts be chosen based on "practical experience in implementing, managing, and evaluating [the desired expertise topic]; research experience in studying [the desired expertise topic]; publications on the topic, and so on" [44]. Based on these recommendations, I chose recruits for the Delphi questionnaires who have demonstrated experience in managing and evaluating file formats in a digital preservation environment, conducting research on file formats in digital preservation, and/or producing publications on the topic. These people have demonstrated experience in these areas either through producing publications, giving presentations, teaching workshops or courses, or writing blog posts about working with or evaluating file formats in a digital preservation context. Additionally, several people were identified as file format experts by experts already identified for the study.

Delbecq, Van de Ven, and Gustafson [45] recommended that for a homogenous group, ten to fifteen participants is adequate to form a Delphi panel. Accordingly, the aim for this study was to assemble two groups of 10-15 expert participants for the two-phase Delphi portion of the study. I initially recruited a total of 25 participants for the Delphi studies. Of these twenty-five participants, four dropped out of the study before the Delphi questionnaire process began. Twenty-one participants completed all or most of the Delphi questionnaires, with 10 participants in one study, and 11 in another.

Participants reported file format experience ranging from one to thirty years. The twenty-one participants reported a total of 210 years of working with file formats in a digital preservation context. The study includes some participants with a comparatively low number of years of relevant experience, but who are included because of the high quality of experience reported.

I recruited one additional participant to serve as a special reviewer for the fourth questionnaire of the study. This reviewer demonstrated a basic understanding of file formats and the challenges they pose to digital preservation. The reviewer demonstrated an aptitude to be trained for this study and was able to demonstrate skills in searching for information about file formats and for rating file format endangerment levels. The reviewer was trained in a one-on-one session where I reviewed the factors, the file formats, and the data collection guide that I created for him.

4. RESULTS

I asked expert participants to rate factors for relevancy as a cause of file format endangerment in order to make sense of the dozens of factors discussed in the literature and to elicit their views on which of the factors have a direct effect on the ability to access information encoded within a particular file format. Both the numerical ratings and participant comments provided insight into this issue.

First, the numerical ratings provided a cutoff for which factors participants believed were at least *somewhat relevant*. With the *somewhat relevant* rating having a value of 0.50, anything that received a rating below 0.50 did not make the cutoff. Half of the factors were rated at 0.50 and above. This cutoff allowed me to eliminate the half of the factors that were rated below 0.50, focusing instead on those factors that the experts deemed to be most relevant. No factor received unanimous ratings of *very relevant*.

Only six factors were rated at 1.00, which is the halfway point between *somewhat relevant* and *very relevant*. If I were selecting factors based solely on the data collected from this Delphi study, this would be the most logical cutoff point, as 1.00 is a good candidate value for a simply "relevant" rating. The factors that were rated at 1.00 and above were: *specification quality* (1.00), *expertise available* (1.05), *community/3rd party support* (1.05), *technical dependencies* (1.05), *rendering software available* (1.14), and s*pecifications available* (1.41).

The comments from participants provided insight into the complex nature of the issue. Many of the comments reflected the ambiguity of some of the factors. For example, one participant wrote about *complexity*, "This is an 'it depends' answer - complexity is hard to bundle into one type of characteristic. Different types of complexity could be answered on their own." Another wrote on the *cost* factor, "I agree with round 1 responses that state cost as a complex, multi-faceted and organizational[ly] influenced factor." Other factors proved to be less ambiguous and participants were able to more directly justify their ratings.

The fact that only six factors were rated at 1.00 and above is an important finding. I began this research with a total of 138 individual factors that I found in the literature. I was able to reduce this list of factors to 21 factors. Through the Delphi process, I was then able to reduce this number to six factors that participants rated as at least halfway between *somewhat relevant* and *very relevant*. Reducing the number of factors this amount was a large step toward the final selection of clear formative indicators for a file format endangerment index.

Table 2 shows a comparison of ranked factors in order of prevalence (in the case of the format rating justification text count) and rating level (Delphi factor rating means and special rater ratings). Examining each dataset included in this table reveals cutoff points for which factors are the most important for indicating file format endangerment. In the Delphi format rating justification text coding count data, there was a distinct drop-off of factor appearances after *specifications available*. While *legal restrictions* appeared in the format rating justification text 97 times, the next most frequently appearing factor, *complexity*, only appeared 63 times. This left *rendering software available*, *ubiquity*, *specifications available*, and *legal restrictions* as well-agreed-upon factors to consider in further analysis.

A logical cutoff point for both the Delphi factor rating mean ranking and special rater factor ratings datasets is a rating above 1.00, the halfway point between *somewhat relevant* and *very* *relevant*. A rating above 1.00 indicates that the factor was rated close to *very relevant*, whereas factors rated at or below 1.00 are at most *relevant*. For the Delphi factor rating mean ranking this leaves the factors *specifications available*, *rendering software available*, *technical dependencies*, and *community/3rd party support*. For the special rater factor ratings this leaves *rendering software available*, *specifications available*, *ubiquity*, and *community/3rd party support*.

Table 2. Factor data comparison chart demonstrating cutoff points for emergent and most relevant factors

Delphi Format Rating Justification Text Factors (# of appearances in text)	Delphi Factor Rating Mean Ranking (mean rating value)	Special Rater Factor Ratings (mean rating value)
Rendering Software Available (162)	Specifications Available (1.40)	Rendering Software Available (1.50)
Ubiquity (130)	Rendering Software Available (1.10)	Specifications Available (1.50)
Specifications Available (111)	Technical Dependencies (1.10)	Ubiquity (1.50)
Legal Restrictions (97)	Community/3rd Party Support (1.10)	Community/3rd Party Support (1.50)
Complexity (63)	Expertise Available (1.00)	Legal Restrictions (0.50)
Community/3rd Party Support (51)	Legal Restrictions (1.00)	Technical Dependencies (0.50)

After comparing the results from the three sets of collected data, five factors emerged as being either more highly ranked, or as appearing more times in the format-rating justification text. Examining each of the five remaining factors in light of the qualitative data collected provides more clarity for which are the most relevant as candidate causal indicators of file format endangerment.

Rendering software available. *Rendering software available* and *specifications available* are the only two factors that appeared beyond the cutoff point in all three datasets. It appeared as the top factor in two of the three datasets, and would have tied for the top ranking in the Delphi factor rating dataset if not for one *not relevant at all* rating. The rationale for this aberrant rating was justified that the participant considered the lack of rendering software to be the definition of obsolescence/file format endangerment and therefore rated it as being not relevant within the context of the participant's self-selected definition.

Four of the eight participants who rated this factor as *very relevant* indicated lack of rendering software strongly suggests file format obsolescence. For example, one participant wrote, "By definition without rendering software the format is obsolete." By far, the comments about the *rendering software* factor in the Delphi factor rating exercise were very strong, simple, and direct: without rendering software a file format is essentially obsolete. The strength of the comments about this factor points to it being a very strong candidate as a direct cause of file format endangerment.

Specifications available. Like *rendering software available*, the *specifications available* factor was included beyond the cutoff point in all three factor evaluation datasets in this study. It received a very high relevancy rating (1.40 of 1.50 possible) from the Delphi factor rating participants. Delphi participants indicated that having specifications available enables the creation of rendering software if none is available. Furthermore, others indicated that it helps to determine if software faithfully renders the contents of a file. One participant wrote, "It is hard to see that a format would not be more endangered if specifications could not be obtained." Based on the ratings and the strength of the participant comments, the *specifications available* factor is another strong candidate as a cause of file format endangerment.

Ubiquity. The case for considering the *ubiquity* factor as a cause of file format endangerment is weakened for several reasons. First is the fact that it only remained above the cutoff point in two of the three datasets. Second, though the special rater rated it as *very relevant*, he explained later that he only considered it to be a secondary factor, because of the following scenario: "there are also formats that are not widely distributed that are not endangered at all, such as the .nes format, used for ROM dumps of Nintendo Entertainment System cartridges."

This sentiment is echoed in many of the Delphi factor rating comments, where several participants described its effect on endangerment in secondary terms. For example, one participant wrote, "The popularity of a given file format increases the support provided by user communities and consequently increases the resources allocated/available for development/maintenance for further developments." In this scenario, the ubiquity of the file format has an effect on other factors that directly affect the endangerment level of the format and serves more as a tertiary factor that affects *community/3rd party support*.

Community/3rd party support. This factor is ultimately a secondary factor, even though it appeared above the cutoff point in two of the three datasets. Participants in the factor rating Delphi referred to it as a stopgap against a single point of failure: "single-point of failures are serious potential problems, and having a format which is supported by a single provider, rather enjoying larger community and 3rd party support, is a classic single point of failure situation. The wider the experience with and understanding of a format, the better, and the lack of those can present serious risks." In this case, community/3rd party support is a factor that can directly support the existence of rendering software, but is often contingent on the availability of specifications.

Technical dependencies. This factor appeared above the cutoff line in only the Delphi factor rating dataset. The special rater noted that he "didn't find technical dependencies to be a useful indicator as all formats have some technical dependencies." When the format rating Delphi participants mentioned technical dependencies, it was typically in the context of causing problems with the full and faithful rendering of a file that calls in information from external files; but do not mention it preventing a file from being rendered at all. In this case, *technical dependencies* is a tertiary factor where *rendering software* is the primary and *rendering software feature/functionality/behavior support* is the secondary factor.

Legal restrictions. This factor appeared above the cutoff line in only the Delphi format justification text coding dataset. Close examination of temporal priority reveals that while legal restrictions do have an effect on accessibility of digital content, this factor is actually a secondary factor to *specifications available* and *community/3rd party support*. The instances where legal restrictions were coded in the format rating justification text were those times where participants mentioned the availability of specifications and the existence of open source software. Legal restrictions can prohibit the free availability of specifications and prohibits the creation of rendering software through third parties.

It was through the process of comparing these results and scrutinizing the remaining factors that I was able to make a final reduction in factors from six to three: *rendering software available*, *specifications available*, and *community/3rd party support.* From beginning to end, I was able to reduce the list of factors from the original 138 factors that I found in the literature to three, for a total reduction of 135 factors.

5. CONTRIBUTIONS AND IMPLICATIONS

The findings of this research suggest that the three top contenders for use in a file format endangerment index are *rendering software available*, *specifications available*, and *community/3rd party support*. This is a marked reduction from previous total of the 21 factors synthesized from the original list of 138 factors I found in the literature. The benefit of which is that file format data collection can be focused in the areas defined in the index.

The research discussed here is the first step toward creating a file format endangerment index that can be used to detect when content encoded in a particular file format may be more difficult to access over time. Following the recommendations of Diamantopoulos and Winklehofer [23] for constructing an index, the next steps are to test and validate the index. Testing and validating an index first requires that data be collected for the selected formative indicators.

A starting point for data collection can be to use the data collected by the special rater and the data collection suggestions provided by the factor rating Delphi participants. From there the index can be validated against the file format ratings collected in the format rating Delphi study, future collected expert ratings, and other external sources. From there, continued data collection for each of the factors can be conducted in conjunction with continued assessment of the collected data.

Once the factors selected for the index have been adjusted and validated, the measure can be put to use in evaluating file format endangerment levels both in the local and global contexts. Coordination of cooperative efforts with institutions, coalitions, and other researchers who are working in this area can expand data collection and the application of the index.

Additionally, it will be valuable to explore nuances of each of the factors. For example, the factor, *specifications available*, could be examined not just by whether or not specifications are available, but by how useful the specifications are to the creation or recreation of viable rendering software. Additionally, the factor, *rendering software available*, could be evaluated not just for whether or not software is available, but how faithfully it represents the original intended representation of the encoded content.

In performing this study, I have used a hypothetical Occam's Razor to cut away what had previously been an unmanageably large collection of mostly inoperable file format endangerment factors to leave just three factors that can be used in a file format endangerment index. The simplification of factors and the creation of the file format endangerment index contributes to the digital preservation community's ability to know which file formats are at risk so issues can be addressed before they becomes too expensive and time consuming to manage in the future.

6. ACKNOWLEDGMENTS

I would like to thank all of the study participants. They generously donated a great deal of time to this study and provided surprisingly detailed and thoughtful responses to the questionnaires. I would also like to thank my dissertation chair and dissertation committee for their guidance and support in the conception and execution of this research.

7. REFERENCES

[1] Merriam-Webster. 1994. *Merriam-Webster's collegiate dictionary* (10th ed.). Springfield, MA: Merriam-Webster, Inc.

[2] van der Knijff, J. 2013a, September 30. *Assessing file format risks: searching for Bigfoot?* Message posted to Open Planets Foundation blogs at http://www.openplanetsfoundation.org/blogs/2013-09-30-assessing-file-format-risks-searching-bigfoot

[3] van der Knijff, J. 2013b, October 8. *Measuring Bigfoot*. Message posted to Open Planets Foundation blogshttp://www.openplanetsfoundation.org/blogs/2013-10-08-measuring-bigfoot

[4] Curtis, J. 2006. *AONS system documentation* (Revision 169 2006-09-29). Technical Report. Australian Partnership for Sustainable Repositories.

[5] Pearson, D., & Webb, C. 2008. Defining file format obsolescence: A risky journey. *International Journal of Digital Curation,* 3,1, 89-106.

[6] Anderson, R., Frost, H., Hoebelheinrich, N., & Johnson, K. 2005. The AIHT at Stanford University. *D-Lib Magazine*, 11,12.

[7] Arms, C.R., & Fleischhauer, C. 2005. Digital formats: Factors for sustainability, functionality, and quality. *Imaging Science & Technology Archiving 2005*, Washington, DC, (April 2005), 222-227.

[8] Becker, C., & Rauber, A. 2011. Decision criteria in digital preservation: What to measure and how. *Journal of the American Society for Information Science and Technology*, 62, 6, 1009-1028.

[9] Scalable Preservation Environments [SCAPE]. n.a. *About SCAPE*. Retrieved December 29, 2013 from http://www.scape-project.eu/about

[10] Faria, L. 2013. *Scout - A preservation watch system*. Retrieved December 29, 2013 from the Open Planets Foundation website http://www.openplanetsfoundation.org/blogs/2013-12-16-scout-preservation-watch-system

[11] Faria, L., Akbik, A., Sierman, B., Ras, M., Ferreira, M., & Ramalho, J.C. 2013. Automatic preservation watch using extraction on the web. *Proceedings of the10th International Conference on the Preservation of Digital Objects, Lisbon, Portugal*.

[12] Graf, R. & Gordea, S. 2013. A risk analysis of file formats for preservation planning. *Proceedings of the10th International Conference on the Preservation of Digital Objects, Lisbon, Portugal*.

[13] Graf, R., Gordea, S., & Ryan, H. 2014. A model for format endangerment analysis using fuzzy logic. *Proceedings of the11th International Conference on the Preservation of Digital Objects (iPres2014) (accepted for publication), Melbourne, Australia*.

[14] Hunter, J. & Choudhury, S. 2006. PANIC: an integrated approach to the preservation of composite digital objects using semantic web services. *International Journal of Digital Libraries*, 6,2, 174-183.

[15] Hunter, J. & Choudhury, S. 2004. A semi-automated digital preservation system based on semantic web services. In *Global Reach and Diverse Impact: Fourth ACM/IEEE-CS Joint Conference on Digital Libraries* (JCDL '04), June 7-11, 2011,Tucson, AZ, pp. 268-278. Association for Computing Machinery.

[16] Ex Libris Group. 2010. *Ex Libris Rosetta: A digital preservation system product description.* Retrieved February 21, 2013 from http://www.exlibrisgroup.com/category/RosettaOverview

[17] Peled, I. 2011. The challenges of building Ex Libris Rosetta, a digital preservation system. *Liber Quarterly*, 21, 1. http://liber.library.uu.nl/index.php/lq/article/view/8012/8354

[18] Tilbury, J. 2014. *The active preservation of digital information*. White Paper. Tessella Group.

[19] Oltmans, E., van Diessen, R.J., & van Wijngaarden, H. 2004. Preservation functionality in a digital archive. In *JCDL 2004: Proceedings of the Fourth Acm/Ieee Joint Conference on Digital Libraries: Global Reach and Diverse Impact: Tucson, Arizona, June 7-11, 2004,* edited by Hsinchun Chen, Michael Christel and Ee-Peng Lim, 279-86. New York, NY: ACM Press.

[20] Koninklijke Bibliotheek. n.d.. *More about the e-Depot*. Retrieved June 7, 2013 from http://www.kb.nl/en/expertise/e-depot-and-digital-preservation/more-about-the-e-depot.

[21] Diamantopoulos, A., Riefler, P., & Roth, K.P. 2008. Advancing formative measurement models. *Journal of Business Research,* 61, 1203-1218. P. 1205

[22] Bollen, K. A. 1989. *Structural equations with latent variables*. New York: Wiley-Interscience.

[23] Diamantopoulos, A. & Winklhofer, H.M. 2001. Index construction with formative indicators: An alternative to scale development. *Journal of Marketing Research*, 38(2).

[24] Petter, S., Straub, D., & Rai, A. 2007. Specifying formative constructs in information systems research. *MIS Quarterly, 31*,4, 623-656.

[25] Lawrence, G.W., Kehoe, W.R., Rieger, O.Y., Walters, W.H., & Kenney, A.R. 2000. *Risk management of digital information: A File format investigation*. Washington, DC: Council on Library and Information Resources.

[26] Fischer, T. 2003. LaTeX as an archiving format: Benefits and problems. *Proceedings of the Sixth International Symposium on Electronic Theses and Dissertations ETD2003*. Berlin: Humboldt-Universität zu.

[27] Huc, C., et al. 2004. *Criteria for evaluating data formats in terms of their suitability for ensuring information long term preservation*. Technical Report. Groupe Pérennisation des Informations Numériques.

[28] Clausen, L.R. 2004. *Handling file formats*. Technical Report. Kongelige Bibliotek.

[29] Christensen, S. 2004. *Archival data format requirements*. Technical Report. Kongelige Bibliotek.

[30] Stanescu, A. 2004. Assessing the durability of formats in a digital preservation environment: The INFORM methodology. *D-Lib Magazine,* 10,11.

[31] Anderson, R., Frost, H., Hoebelheinrich, N., & Johnson, K. 2005. The AIHT at Stanford University. *D-Lib Magazine, 11*(12).

[32] Shirky, C. 2005. *Library of Congress Archive Ingest and Handling Test (AIHT): Final report*. Technical Report. National Digital Information Infrastructure & Preservation Program (NDIIPP).

[33] Cornwell Management Consultants. 2005. *Selection of preservation formats: trends and issues*. Technical Report. The National Archives, U.K.

[34] Cornwell Management Consultants. 2005. *Criteria for the selection of preservation formats*. Technical Report. The National Archives, U.K.

[35] Ferreira, M., Baptista, A.A., & Ramalho, J. C. 2006. A foundation for automatic digital preservation. *Ariadne, 48*.

[36] Ferreira, M., Baptista, A. A., & Ramalho, J. C. 2007. An intelligent decision support system for digital preservation. *International Journal on Digital Libraries*, 6, 4, 295-304.

[37] InterPARES. 2007. General study 11 final report: Selecting digital file formats for long-term preservation (Version 1.1). British Columbia, Canada: McLellan, E. P.

[38] Barve, S. 2007. File formats in digital preservation. In Madalli, D.P, & Madalli, P. (Eds.), *International Conference on Semantic Web & Digital Libraries: ICSD-2007*, 239-248.

[39] Rog, J., & Wijk, C, van. 2008. *Evaluating file formats for long-term preservation*. Technical Report. Koninklijke Bibliotheek.

[40] Becker, C., Kulovitz, H. Brown, A. 2008. *Planets: Report on service integration in Plato 2*. Technical Report. Planets Project.

[41] Helmer, O., & Rescher, N. 1959. On the epistemology of the inexact sciences. *Management Science*, 6,1, 25-52.

[42] Dalkey, N.C. 1968. Predicting the future. *National Conference on Fluid Power*, Chicago, Illinois.

[43] Gordon, T. J., & Helmer, O. 1964. *Report on a long-range forecasting study*. Technical Report. RAND Corporation.

[44] Luo, L., & Wildemuth, B. M. 2009. “Delphi studies.” In B. M. Wildemuth (Ed.), *Applications of social research methods to questions in information and library science*. Westport, CT: Libraries Unlimited.

[45] Delbecq, A.L., Van de Ven, A.H., & Gustafson, D.H. 1975. *Group techniques for program planning*. Glenview, IL: Scott, Foresmann, and Co.

The Dendro research data management platform

Applying ontologies to long-term preservation in a collaborative environment

João Rocha da Silva
Faculdade de Engenharia da
Universidade do Porto /
INESC TEC
Portugal
joaorosilva@gmail.com

João Aguiar Castro
Faculdade de Engenharia da
Universidade do Porto /
INESC TEC
Portugal
joaoaguiarcastro@gmail.com

Cristina Ribeiro
DEI — Faculdade de
Engenharia da Universidade
do Porto / INESC TEC
Portugal
mcr@fe.up.pt

João Correia Lopes
DEI — Faculdade de
Engenharia da Universidade
do Porto / INESC TEC
Portugal
jlopes@fe.up.pt

ABSTRACT

It has been shown that data management should start as early as possible in the research workflow to minimize the risks of data loss. Given the large numbers of datasets produced every day, curators may be unable to describe them all, so researchers should take an active part in the process. However, since they are not data management experts, they must be provided with user-friendly but powerful tools to capture the context information necessary for others to interpret and reuse their datasets. In this paper, we present Dendro, a fully ontology-based collaborative platform for research data management. Its graph data model innovates in the sense that it allows domain-specific lightweight ontologies to be used in resource description, acting as a staging area for later deposit in long-term preservation solutions.

Categories and Subject Descriptors

H.3 [**Information Search and Retrieval**]: Online Information Services Data sharing

Keywords

Research data management, data curation, ontologies, data repositories, Dendro

1. INTRODUCTION

It is widely accepted that research data management should start as soon as possible in the research workflow. However, most research data management platforms like CKAN, Zenodo or Dryad are designed for publishing “finished” datasets that can be *cited*. This *a posteriori* data management timing yields very high-quality and highly-selected datasets, but in many cases the number of datasets that are actually published can be quite low. Empty dataset archives and repositories are still commonplace [Nelson 2009; Borgman 2012].

Several data management projects focus on supporting collaboration within research groups and making daily data management activities easier. The resulting tools are therefore entry points through which the datasets can enter a preservation workflow [Hodson 2011; Shotton 2012]. These solutions focus on providing easy-to-use shared storage spaces with regular automated backups, connected to a data repository. The main objectives were to capture data as early as possible and leave detailed description for later (*curation by addition*). In both cases, only a minimal set of metadata is required upon initial submission, leaving the decision to enrich the metadata to the researcher and/or curator.

Current data management platforms often limit the metadata that can be added to a dataset to generic descriptors (e.g. Dublin Core) or a pre-existent set of descriptors that depositors are asked to fill in at the time of deposit. CKAN [Open Knowledge Foundation 2014] is an exception, as it allows an additional set of arbitrary metadata to be added to deposited datasets, in the form of *ad-hoc* text fields. This allows domain-specific metadata to be recorded, although without any pre-defined meaning or standards-compliance.

Dendro, our proposed research data management platform, aims to establish a tradeoff between close proximity to the researcher, incremental data description, quick and simple deposit and no metadata requirements. It uses a triple store to support an ontology-based data model in order to satisfy the metadata needs of different research communities. No metadata requirements exist at the time of deposit, but the basic descriptors (creator, modification date, creation date, etc.) are provided, and the user is expected to fill them in. Richer descriptors are presented as *recommendations* that researchers and curators can choose to fill in or not for each resource. From a preservation standpoint, it is completely supported by open-source software built for cloud-level scalability. Its underlying data model makes data easier to preserve due to its intrinsic readability and Linked Open Data foundation. Its dispenses a relational database and is designed to foster dataset integration in the Semantic Web as Linked Open Data (LOD). An interesting side-effect that stems from the adoption of this model is that the usual layers of relational-LOD translation logic that often exist in solutions that provide LOD compatibility solutions are eliminated. An practical example is Semantic MediaWiki, that uses a relational database in its transactional system and an RDF store for semantic querying, requiring specific code to maintain a permanent mapping between the two solutions. We argue that, by removing the dependency on a relational database altogether, we can remove the concerns over its migration when the system is rendered obsolete and provide an ontology-based metadata model from end to end.

2. A TRIPLE-BASED DATA MODEL

Unlike key-value metadata representations, a linked data representation gives structure and explicit *meaning* to metadata values, allowing datasets, papers, researchers and other research-related resources to be connected by meaningful links. These meanings can also be reused from existing specifications (*ontologies*) or newly created if no ontology defines them. The advantages of this representation from a preservation point of view include the simplicity of the data model and its superior flexibility (it can grow incrementally as more ontologies for different domains are designed). When registering the URI of the creator's web page in the `dc:creator` of a dataset, a system built on linked data will record the *meaning* of that string value, unlike a relational system, where there is no distinction between different types of values. These meanings are specified with ontologies, which can be shared along with the data and the metadata records. In a preservation environment, the advantages are clear: linked data provides great support for self-documented metadata which can also be represented in RDF format—an open, plain-text representation with minimal reliance on specific processing software.

Dendro was designed from the start as a user-friendly interface targeted at users without data management skills. As they interact with the system, a linked data knowledge base is built using ontologies in the background. It is similar to a semantic wiki in the sense that it allows users to collaboratively shape the underlying graph through their daily interaction and directly uses ontologies for parameterization (no mapping between a relational model and a triple store representation ever occurs). Moreover, Dendro's data model is built to offer programmers the appropriate granularity for descriptor-level analysis, allowing the easy combination of descriptors from several domains. We illustrate this by comparing Dendro's data model with the data model of Semantic MediaWiki, perhaps the most widely known semantic wiki.

2.1 Dendro vs. Semantic MediaWiki

Semantic MediaWiki (SMW) is built around ontologies that are used to give *meaning* to the links established between wiki pages (*semantic links*). It offers two different interfaces for establishing semantics between wiki pages. The first one is the standard text editor where semantics can be added to a link tag. For example, one can write: `The author of this paper was [[author:Bob].` In a wiki page. The result would be a very small wiki page with link to **Bob**'s page in the wiki. Internally, a link would be established between the page being edited and the web page of the author. To apply this technique to dataset description, one would start by creating a wiki page for each file in a dataset and write a plain text description containing several of these links. This way, semantic metadata could be embedded in the metadata descriptions.

Another alternative is using SMW's *semantic forms.* These are more structured interfaces designed for users to fill in a predefined set of links. However, these predefinitions have to be specified *a priori*; researchers cannot select descriptors to include in their metadata sheets, having instead to rely on a single template.

Our past work on DataNotes, an extension to SMW [Rocha da Silva et al. 2013] proposed a modification to the platform to allow researchers to freely include descriptors from several ontologies in their descriptions. Extensive changes had to be made to the business logic and user interface, but the issues caused by having a relational and a triple-based side by side still remained.

2.2 The advantages of a graph-based model

Ontologies and triple stores allow us to tackle the research data management challenge in a unique manner, enabling the representation of resources with different sets of attributes, even when they are not known at the time of modeling. Realizing the advantages of a graph-based data model over the constraints posed by a relational approach, a design for a multi-domain research data management system has proposed a similar ontology-based architecture built on triple stores [Li et al. 2013].

The data model behind Dendro has the right granularity for describing any kind of resource using variable descriptors without incurring in a convoluted relational database schema, which would mean complex queries and heavy JOIN operations every time we wanted to access the descriptors of a resource. Also, since the core data model of the platform uses a triple store, it becomes possible to directly load ontologies from different domains into the knowledge base and reuse the concepts specified in those ontologies. This allows domain experts to specify their own ontology using high-level tools like Protégé[1] (or just reuse existing ones) and load them into Dendro, thus enabling the new concepts to be used in the description of research data assets. Given the open nature of ontologies and their asynchronous evolution through reuse, platforms like Dendro can retain a higher level of interoperability than conventional RDB-based ones. With this approach we plan for obsolescence in a positive way: the data more easily survive the obsolescence of the Dendro platform, as the contents of the entire data model can be exported as Linked Open Data (LOD). The data model itself will also be public and self-documented, since

[1] `http://protege.stanford.edu`

it is good practice of ontology design to document ontology concepts at design time, via the common `rdfs:label` and `rdfs:comment` description properties—information that is also used by Dendro in its user interfaces.

2.3 Dendro in the preservation workflow

Figure 2 shows Dendro's role in the research data management ecosystem as it supports the process at different points in time.

1. Data creation, description and sharing within the research group throughout their research activities (**1**). Dendro provides a friendly web interface for humans as well as a series of APIs to enable other systems to manipulate files and folders as well as their metadata. Metadata creation is carried out using properties from different ontologies (either already present on the web or modeled by curators). With a triple store as the storage and querying layer, metadata can be added as property instances. Resources can also be retrieved using SPARQL queries, making faceted searches much easier to implement than on a relational model. Moreover, the simple triple store model enables external entities to easily query the data store via SPARQL.

2. Dataset deposit, where a set of files from Dendro, as well as their relevant metadata, are packaged and deposited in a long-term preservation platform such as Zenodo or CKAN(**2**)

3. Evolution of metadata recommendations (**3**). As the metadata specifications for different domains are created, they are also shared on the web, encouraging reuse and community-driven maintenance. Descriptor semantics become publicly documented and available for reuse in other data management systems, enabling a continuous evolution process that contributes towards the emergence of some ontologies as metadata standards for different research domains.

4. Data reuse (**4**). When a researcher accesses a dataset, documentation on the meaning of each descriptor will be available in the ontology from where that descriptor originated, making the interpretation of domain-specific metadata easier.

3. OVERVIEW OF THE PLATFORM

When designing the tools for an integrated preservation environment, one must ensure that the data stored within can survive the obsolescence of the environment itself. Dendro's triple-based data model, its reliance on shareable ontologies and a full open-source technology stack all contribute to maintaining access and interpretation of the stored datasets even after the platform's decommissioning.

Figure 1 shows the architecture of Dendro. The "Data" layer holds the data model for the platform, composed of three subsystems: an OpenLink Virtuoso Database (Open-Source version), an ElasticSearch server to enable distributed document indexing and a MongoDB/GridFS file storage cluster. The graph database is used to represent all the resources in the knowledge base (for example, `Researchers`, `Files`, `Folders` and their attributes, represented using existing ontologies. Some of the ontologies being used at this time are Dublin Core Terms Ontology (for all resources in general), the Nepomuk File Ontology (for files and folder structures representation) and the Friend of a Friend Ontology (for describing platform `Users`). All queries specified by the *Logic* layer are sent to OpenLink Virtuoso's SPARQL endpoint. In case Virtuoso becomes obsolete, Dendro's triple-based model is designed to live on, since it can be fully exported in RDF and imported into another RDF-compliant solution. The triples plus the ontologies made available on the web enable a complete understanding of the stored information.

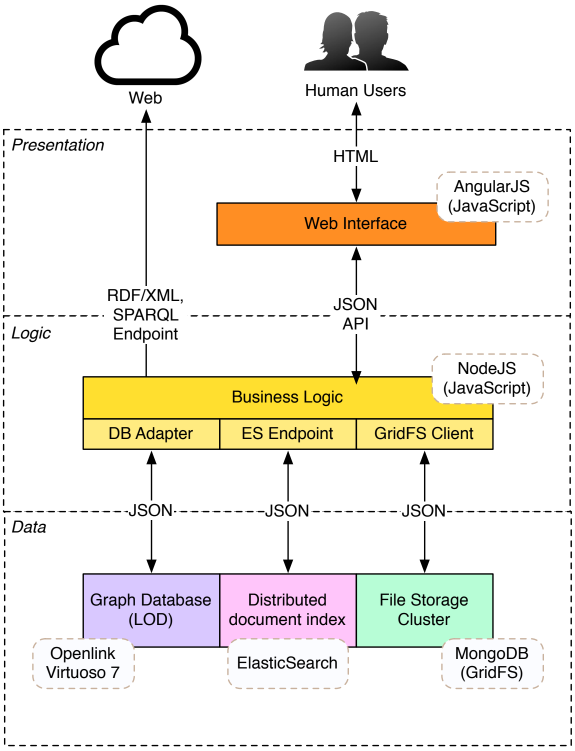


Figure 1: Dendro's architecture and technology stack

The *Logic* layer comprises Dendro's business logic, and includes three endpoints that connect to the underlying Data layer. A Database Adapter was written from scratch in order to provide a higher level of abstraction over the REST API provided by OpenLink Virtuoso. The module automatically performs the conversion between the results format provided by Virtuoso and Javascript objects to provide programmers an abstraction over the database, similar to Hibernate for Java or LINQ in the .NET platform.

The Logic Layer is written in NodeJS for handling large numbers of simultaneous connections—this allows numerous users or external systems (via Dendro's API) to interact directly with the platform to manage data and metadata. Dendro is primarily written in JavaScript, a simple and very widely known and used programming language among web developers—a plus when planning for an open-source preservation effort, as a large potential developer base makes main-

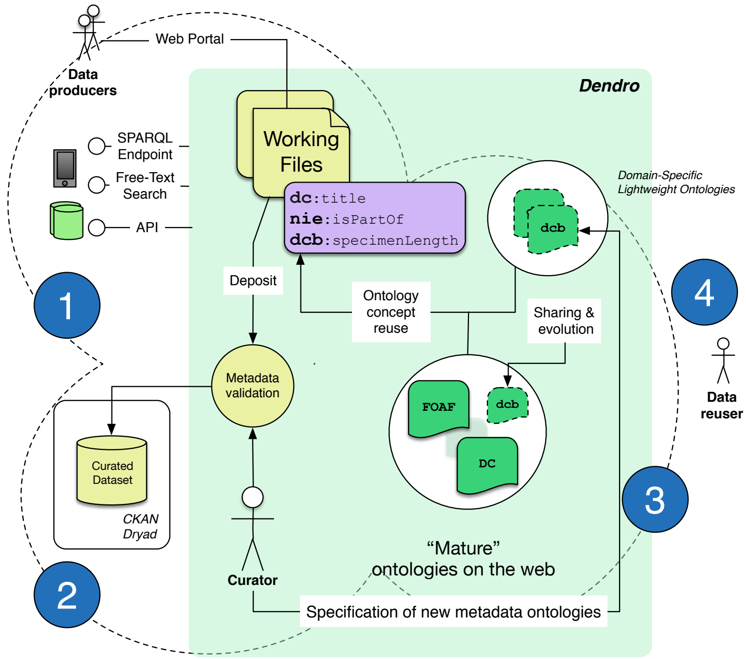


Figure 2: Dendro's role in a research data management ecosystem

tenance and evolution easier.

4. MANAGING DATASETS USING DENDRO

Figure 3 shows Dendro's main interface and the representation of recorded metadata in the triple store. Area **1A** shows the operations that can be performed over the current folder: Create a new folder, upload files, download the current folder, backup the current folder (includes metadata), restore a folder from a backup, and hide deleted files.

Area **1B** is the file explorer, showing the contents of the currently open folder. **1C** is a search box that allows any resource to be retrieved by any *literal* value (a continuously-updated index powered by ElasticSearch). **1D** exemplifies how domain-specific descriptors can be added to a metadata description; in this case, the `SpecimenLength` descriptor is added to the metadata for this folder. This descriptor has been previously specified in an ontology designed for mechanical engineering. Other descriptors from different ontologies can be loaded into the system, and the autocomplete box will retrieve them based on the values of their `rdfs:label` and `rdfs:comment` description properties. When a descriptor is selected by the user, it is added to the metadata editing area of the interface in the center. At the same time, the ontology from which it originates is "locked" so that the interface will suggest additional descriptors from the same ontology in a *quick-access* list of descriptors (Area **1E**). When a metadata value is inserted, it is recorded in the underlying triple store.

Area **2** shows a simple SPARQL query that obtains all the properties that have the folder being described as their subject. Although this is a very simple example, SPARQL allows resources in the knowledge base to be easily retrieved based on their properties and also on the properties of their linked resources. The results of the query are shown in (**3**)—note the descriptors from three different ontologies: *Dublin Core* (for generic metadata), *Nepomuk Information Element* (for file-related information) and *Double Cantilever Beam*, the domain-specific ontology for fracture mechanics datasets.

5. CONCLUSIONS

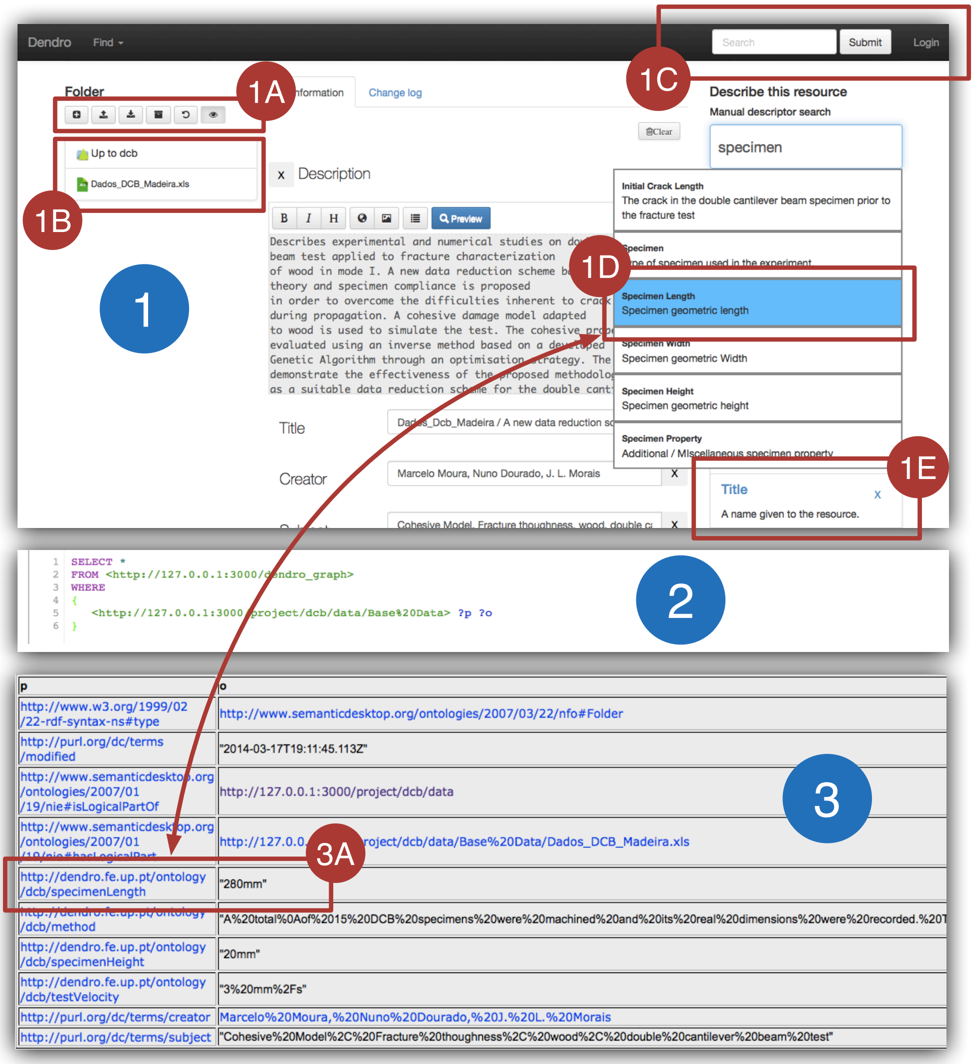
In this paper, we have presented Dendro, a collaborative research data management platform built on a triple store data model. Comparing it with repository platforms built on relational databases, we can see that the fully ontology-based data model provides a much more preservation-friendly environment, as it becomes self-documented. The *meaning* of the metadata values is specified in ontologies, which can evolve asynchronously according to the needs of different domains and be shared and retrieved from the web.

By representing datasets, papers, researchers and other research assets as resources and dataset metadata as values for properties relating these resources, a simple (triple-based) extensible (via ontologies) and powerful (supporting SPARQL querying) data model can be built.

Preliminary studies show that the platform satisfies several data management capabilities requested by researchers in our previous studies. We are now working on improving and testing it with researchers from different domains, while improving its interaction with existing repository platforms.

6. ACKNOWLEDGEMENTS

This work is supported by project NORTE-07-0124-FEDER-000059, financed by the North Portugal Regional Operational Programme (ON.2–O Novo Norte), under the Na-



x Description

Describes experimental and numerical studies on do
beam test applied to fracture characterization
of wood in mode I. A new data reduction scheme b
theory and specimen compliance is proposed
in order to overcome the difficulties inherent to crack
during propagation. A cohesive damage model adapted
to wood is used to simulate the test. The cohesive prop
evaluated using an inverse method based on a developed
Genetic Algorithm through an optimisation strategy. The
demonstrate the effectiveness of the proposed methodolo
as a suitable data reduction scheme for the double cant

Title: Dados_Dcb_Madeira / A new data reduction sc

Creator: Marcelo Moura, Nuno Dourado, J. L. Morais X

Cohesive Model, Fracture thoughness, wood, double ca X

Initial Crack Length
The crack in the double cantilever beam specimen prior to the fracture test

Specimen
ype of specimen used in the experiment.

Specimen Length
Specimen geometric length

Specimen Width
Specimen geometric Width

Specimen Height
Specimen geometric height

Specimen Property
Additional / MIscellaneous specimen property

Title x
A name given to the resource.

```
SELECT *
FROM <http://127.0.0.1:3000/dendro_graph>
WHERE
{
   <http://127.0.0.1:3000/project/dcb/data/Base%20Data> ?p ?o
}
```

p	o
http://www.w3.org/1999/02 /22-rdf-syntax-ns#type	http://www.semanticdesktop.org/ontologies/2007/03/22/nfo#Folder
http://purl.org/dc/terms /modified	"2014-03-17T19:11:45.113Z"
http://www.semanticdesktop.org /ontologies/2007/01 /19/nie#isLogicalPartOf	http://127.0.0.1:3000/project/dcb/data
http://www.semanticdesktop.org /ontologies/2007/01 /19/nie#hasLogicalPart	http://127.0.0. roject/dcb/data/Base%20Data/Dados_DCB_Madeira.xls
http://dendro.fe.up.pt/ontology /dcb/specimenLength	"280mm"
http://dendro.fe.up.pt/ontology /dcb/method	"A%20total%0Aof%2015%20DCB%20specimens%20were%20machined%20and%20its%20real%20dimensions%20were%20recorded.%20T
http://dendro.fe.up.pt/ontology /dcb/specimenHeight	"20mm"
http://dendro.fe.up.pt/ontology /dcb/testVelocity	"3%20mm%2Fs"
http://purl.org/dc/terms/creator	Marcelo%20Moura,%20Nuno%20Dourado,%20J.%20L.%20Morais
http://purl.org/dc/terms/subject	"Cohesive%20Model%2C%20Fracture%20thoughness%2C%20wood%2C%20double%20cantilever%20beam%20test"

Figure 3: Dendro and its interaction with the triple store

tional Strategic Reference Framework (NSRF), through the European Regional Development Fund (ERDF), and by national funds, through the Portuguese funding agency, Fundação para a Ciência e a Tecnologia (FCT). João Rocha da Silva is also supported by grant SFRH/BD/77092/2011, provided by the Portuguese funding agency, Fundação para a Ciência e a Tecnologia (FCT).

References

Christine L. Borgman. 2012. The conundrum of sharing research data. *Journal of the American Society for Information Science and Technology* 63, 6 (2012). http://ssrn.com/paper=1869155http://onlinelibrary.wiley.com/doi/10.1002/asi.22634/full

Simon Hodson. 2011. *ADMIRAL: A Data Management Infrastructure for Research Activities in the Life sciences.* Technical Report. University of Oxford.

Yuan-fang Li, Gavin Kennedy, Faith Ngoran, Philip Wu, and Jane Hunter. 2013. An Ontology-centric Architecture for Extensible Scientific Data Management Systems. *Future Generation Computer Systems* 29, 2 (2013), 1–38.

Bryn Nelson. 2009. Data Sharing : Empty archives. *Nature* 461, September (2009). http://europepmc.org/abstract/MED/19741679

Open Knowledge Foundation. 2014. CKAN documentation - Release 2.2a. (2014).

João Rocha da Silva, José Barbosa, Mariana Gouveia, Cristina Ribeiro, and João Correia Lopes. 2013. UPBox and DataNotes: a collaborative data management environment for the long tail of research data. (2013).

David Shotton. 2012. The JISC UMF DataFlow Project : Introduction to DataStage. *Technical Report* (2012).

A Perspective on Archiving the Scholarly Web

Herbert Van de Sompel
Los Alamos National Laboratory
Los Alamos, NM, USA
http://orcid.org/0000-0002-0715-6126/
herbertv@lanl.gov – @hvdsomp

Andrew Treloar
Australian National Data Service
Melbourne, VIC, Australia
http://orcid.org/0000-0002-8911-3081/
andrew.treloar@ands.org.au – @atreloar

ABSTRACT

As the scholarly communication system evolves to become natively web-based and starts supporting the communication of a wide variety of objects, the manner in which its essential functions – registration, certification, awareness, archiving - are fulfilled co-evolves. This paper focuses on the nature of the archival function based on a perspective of the developing future scholarly communication infrastructure.

General Terms

Infrastructure, preservation strategies and workflows, theory of digital preservation

Keywords

Scholarly communication, web preservation

1. THE FUTURE ACCORDING TO THE PAST

The 2004 paper "Rethinking Scholarly Communication" [3] observed significant trends in the way scholarly communication was then evolving as a result of the gradual, yet steady, transition towards a digital and network-based endeavour.

Based on these observations, the paper revisited the perspective of what constitutes a unit of communication, moving beyond journal publications, and including a wide variety of objects such as datasets, simulations, software as well as compound aggregations of such objects linked together using appropriate relationships.

The paper also pointed at the possibility of a profound reconfiguration of the scholarly communication system enabled by the networked technologies. It did so guided by the theoretical perspective developed by [2] of the essential functions that must be fulfilled by any system of scholarly communication, irrespective of its implementation:

- Registration: Allows claims of precedence for a scholarly finding
- Certification: Establishes validity of claim
- Awareness: Allows actors in the system to remain aware of new claims
- Archiving: Preserves the scholarly record over time

In the system of journals, these functions were vertically integrated in the journal-centric ecosystem: a journal took care of registering claims by accepting manuscripts, of certification through the peer-review process it coordinated, of awareness through its availability in libraries, and of (distributed) archiving by means of its long-term presence on library shelves, worldwide.

However, as soon as the Web made it possible to communicate digital information across a global network, signs of a future in which the functions of scholarly communication would no longer be fulfilled in a vertically integrated manner became apparent:

- The preprint movement, led by arXiv.org, then still at the Los Alamos National Laboratory, demonstrated the value of allowing manuscripts to be submitted (registration) and discovered (awareness) with out applying a certification process to them.
- As soon as journals went digital and were baptized e-journals, their preservation was no longer the sole concern of libraries. Quite to the contrary, publishers and special-purpose organizations such as Portico started fulfilling the archival function, thereby disconnecting it from the tight connection it had for centuries with the awareness function.

The 2004 paper drew these indicators to their logical conclusion by pointing at the future possibility of a web-based scholarly system in which the essential functions are fulfilled in discreet, disaggregated, and distributed manners, and in which a variety of networked pathways interconnect the autonomous hubs that fulfil these functions. Inspired by preprints, and motivated by a desire to increase the speed of discovery, the paper further made a plea in support of early registration – decoupled from certification - of the brave new objects of scholarly communication.

2. THE FUTURE IS NOW

Ten years later, indicators of both the changing nature of the objects of scholarly communication and of the disaggregated fulfilment of the essential functions of a scholarly communication system are abundant. To quote William Gibson, "The future is already here – it's just not very evenly distributed"[1]. Although indicators exist across scholarly disciplines, the life sciences provide the most compelling and complete range of examples, and so will be used to illustrate the ideas presented in this paper:

- Registration: A wide variety of life science objects are being registered in various systems. BioRxiv[2] is a preprint service modelled on arXiv.org. The RCSB Protein Data Bank (PDB)[3] enables the registration of experimentally determined structures of proteins, nucleic acids, and complex assemblies. While autonomous, it does have a tight binding to the journal publication and hence certification process – submissions about such structures will not be accepted without an assigned PDB identifier. WikiPathways[4] provides a

[1] http://www.npr.org/templates/story/story.php?storyId=1067220

[2] http://biorxiv.org

[3] http://pdb.org/pdb/home/home.do

[4] http://wikipathways.org/index.php/WikiPathways

collaborative platform for the registration and curation of biological pathways; because it is based on wiki technology, a version history of all pathways is maintained. NeuroLex[5] is a platform for managing neuroscience terminology; it supports versioning to accommodate terminology evolution over time. NanoPublications[6] are targeted at machine consumption and convey a set of discrete scientific assertions and their provenance expressed as RDF, and are typically obtained by mining journal publications. MyExperiment[7] allows for the registration of scientific workflows.

- Certification: A number of systems exist to enable the community to certify the validity of findings in a manner that is disconnected from the journal's peer-review process. PubMed Commons[8] and PubPeer[9] both allow for post-publication commentary on methods or results. MyExperiment supports certification through social network indicators such as views, downloads, favourites. And, machines are starting to play a role in certification, as exemplified in Project FeederWatch[10] where software detects possible errors in bird species observation/identification data and passes potential errors on to humans for resolution.
- Awareness: Examples of support for the awareness function for novel objects include myExperiment's workflow search engine, and the RSS alerting mechanism used by eLabNotebook[11] to keep researchers informed about experiments as they are conducted.
- Archiving: The archiving function for journals is fulfilled by dedicated services such as Portico[12] and CLOCKSS[13]. For novel objects, dedicated archives exist depending on the content type. For example, the PDB enables the archiving of experimentally determined structures of proteins, nucleic acids. Genbank[14] maintains an annotated collection of all publicly available DNA sequences. While neither of these systems has the long-term commitment of a national archive, they do provide an implied level of ongoing availability.

3. CHARACTERISING THE FUTURE

While [3] anticipated some of the characteristics of a future communication system that have meanwhile emerged, many further characteristics of its ongoing evolution can be observed at this point. These observations pertain both to scholarly communication as such and to the objects that are being communicated; they are summarized in Figure 1 and Figure 2, respectively. In both figures, the left hand side reflects the status of a process or property in the system of journals, whereas the right hand side reflects its status in a future, emerging system, which this paper refers to as the web of objects.

The major observation depicted in Figure 1 is the transition of the research process itself from being hidden in the system of journals towards being visible in the web of objects. Indeed, the increased use of commodity networked technologies such as on demand cloud computing infrastructure and collaboration/sharing platforms for a variety of objects including software and workflows, make sharing objects that are created during that process not only possible but also attractive. MyExperiment, GitHub, Dropbox, networked lab notebooks, scientific wikis and blogs stand out as obvious examples of this.

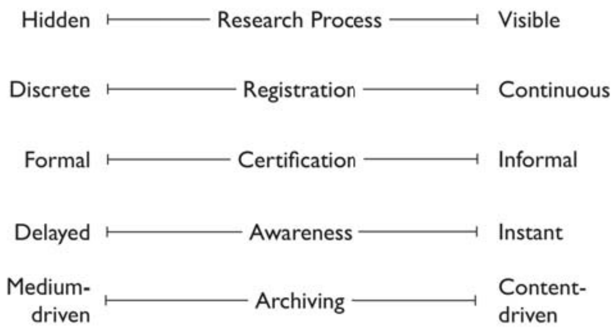


Figure 1: Scholarly Communication Evolution

Figure 1 also considers the evolving nature of the essential functions of a scholarly communication system [2]:

Registration is on a continuum that was characterized by discrete submissions of manuscript to continuous registration of a wide variety of objects, enabled by the aforementioned, networked commodity platforms that are used during the research process.

Certification in the system of journals is conducted in a formal peer-review process, but an evolution towards the inclusion of informal certification approaches, for example, based on indicators extracted from social network interactions is apparent.

Awareness in the system of journals was often delayed by years, in part as the result of lengthy peer-review processes but also due to its – originally – paper based nature. Within the system of journals, a trend towards faster communication can be observed, made possible by electronic distribution but also through an increased focus by certain journals on rapid turn-around peer-review. In the web of objects, awareness is already instantaneous: as soon as an object has a URI, notification technologies (Twitter, RSS, Dropbox alerts, etc.) make immediate discovery possible.

Archiving in the paper-based journal system was characterised by the medium that was being archived. Journals were printed on paper and libraries archived paper irrespective of the content that was printed on it. As the scholarly record evolves to include a variety of objects, including digital journals, an evolution towards content-driven archiving becomes apparent. Indeed, the expertise and infrastructure required to digitally preserve collections of PDF files, discipline-specific datasets, scientific blogs, etc. is significantly different and calls for archival specialization driven by content: Portico archives journal articles, GenBank archives genome sequences, web archives archive web pages, etc.

Figure 2 observes the changing nature of the objects that are communicated in the scholarly communication system, confirming the evolution from fixed to varying, from atomic to compound, from uniform to diverse, and from standalone to inter-related or networked that was anticipated in [3]. In addition, it observes the evolution from journal articles that exhibit a clear sense of fixity towards dynamic objects that (at least during part

[5] http://neurolex.org/

[6] http://nanopub.org/wordpress/

[7] http://mypexperiment.org

[8] http://www.ncbi.nlm.nih.gov/pubmedcommons/

[9]http://pubpeer.com

[10] http://feederwatch.org

[11]http://ourexperiment.org

[12] http://portico.org

[13] http://www.clockss.org/clockss/Home

[14] http://www.ncbi.nlm.nih.gov/genbank/

of their visible life cycle) are continuously changing (for example as they are being collaboratively edited on the aforementioned commodity platforms). The ongoing evolution from restricted to unconstrained access to scholarly objects catalysed by the Open Access and Open Science movements is also depicted.

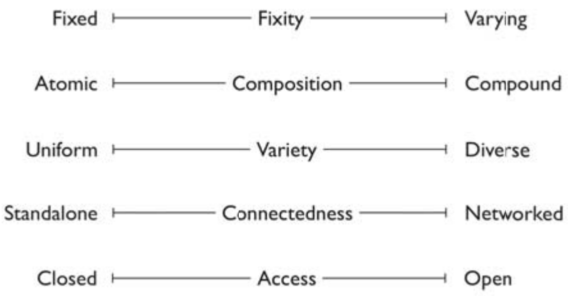


Figure 2: Communicated Object Evolution

4. ARCHIVING THE FUTURE

Several of the aforementioned indicators of the evolution of the scholarly communication system and the communicated objects have a significant impact on the way in which the archival function of a future system can and will be fulfilled. For example:

- In a closed access system, content has to be transported from its original custodian to the designated archive through restricted back office processes. An open system allows for both organized and accidental archiving by means of the open Web, and puts no constraints on the number or kinds of parties that can hold archived copies.
- The suggested evolution from medium-driven to content-driven archives yields an ecosystem of specialized, distributed archives and calls for appropriate levels of cross-archive interoperability in order to support seamless, uniform access to archived objects.

The remainder of this section zooms in on two important areas in which this evolution impacts the archival function of scholarly communication: the increased visibility of the research process and the dynamic, inter-related nature of communicated objects.

4.1 Recording is not archiving

The increased visibility of the research process is, among others, enabled by the adoption of commodity web platforms to record and expose the process. The use of GitHub for the purpose of scientific software development serves as an excellent example of a class of such platforms that share a number of characteristics.

These platforms were not designed with a focus on scholarly use cases, but nevertheless excel at the way in which they fulfil several of the functions of a scholarly communication system. Registration is supported not just by allowing submissions, but also by accurate time stamping and elaborate versioning support. Certification is achieved through a range of reputation-based features such as collaboration, commentary, activity indicators, and likes. Awareness is fulfilled in ways that directly result from the mere presence of these platforms on the open web. This yields discoverability of objects submitted to these platforms through common search engines and the possibility to advertise them on social platforms.

Although these platforms have numerous features that are highly attractive from the perspective of scholarly use cases, it must be observed that they do not fulfil the scholarly archiving function even though their capability to record objects with fine versioning granularity might give the impression they do. Indications that these platforms are excellent recorders of the scholarly process but do not have the long-term commitment to preservation that is expected for objects that are part of the scholarly record can be found in their legal terms and conditions.

Staying with the GitHub example, here are some excerpts from the terms of service[15] that make it explicit that GitHub is not in the archive or persistence business:

> GitHub reserves the right at any time and from time to time to modify or discontinue, temporarily or permanently, the Service (or any part thereof) with or without notice. (E.1)
>
> GitHub does not warrant that (i) the service will meet your specific requirements, (ii) the service will be uninterrupted, timely, secure, or error-free, (iii) the results that may be obtained from the use of the service will be accurate or reliable, (iv) the quality of any products, services, information, or other material purchased or obtained by you through the service will meet your expectations, and (v) any errors in the Service will be corrected. (D.4)

To further clarify the suggested difference between recording and archiving, Table 1 lists some distinguishing characteristics.

Table 1: Recording vs. Archiving

Recording	Archiving
Short-term	Longer-term
No guarantees provided	Attempt to provide guarantees
Write many/read many	Write once/Read many
Scholarly process	Scholarly record

It follows that, as these platforms are increasingly embraced for scholarly use, an appropriate archival function must be overlaid on them to guarantee the long-term integrity of the web based scholarly record. The awareness of this need is growing, as illustrated by the recent announcement[16] of a bridge between GitHub and CERN's Zenodo research output sharing platform, which aims at enabling citation and preservation of code. But, in order to deal with the wide variety of web platforms that is and will be used for scholarship, solutions that connect two distinct environments will not suffice. A systemic solution for the transfer of scholarly objects from the recording platforms into archival environments is required.

4.2 Archiving can not be atomic

The dynamic, compound, and inter-related nature of scholarly communication objects yields significant challenges for the fulfilment of the archival function. In order to illustrate this, consider a comparison between the print era of the system of journals and the web of objects towards which the scholarly communication system evolves.

When published, a journal article references other articles, published in the same or other journals. Both the referencing and the referenced articles are preserved in library stacks, worldwide. In order to revisit the article and its context of referenced articles some time after publication, it suffices to visit the library stacks and pull the appropriate journal issues. Since the content was

[15] http://help.github.com/articles/github-terms-of-service

[16] http://home.web.cern.ch/about/updates/2014/03/tool-developed-cern-makes-software-citation-easier

printed on paper and fixed, the combination of the article and its surrounding context of referenced articles remains the same as it was on the day of the article's publication. Gathering all articles may require some library hopping, but the original information bundle can accurately be recreated.

Reconsider this scenario for the web of objects. Starting conservatively from the same point of a journal article, rather than another scholarly object such as software, still serves as a sufficient illustration. The web-based article not only references other articles but also links to a variety of other objects that reside on the web, such as software, data, project web sites, scientific blogs, etc. Recreating the information bundle made up of the article and its surrounding context some time after publication in this scenario is far less trivial as a result of the dynamic nature of the web, the malleability of content inherent to digital media, and the dynamic nature of scholarly objects especially the ones created in the course of the research process. Indeed, even the links in the article are subject to reference rot, a term coined in the Hiberlink project[17] to refer to the combination of link rot, also known as 404 Not Found, and content drift, the evolution of a web resource's content away from what it was at the moment it was linked, possibly up to a point that it becomes unrepresentative of the content intended by the link. And, while referenced articles themselves may still be frozen in time, they are increasingly embedded in web environments with dynamic content such as commentary, metrics, etc.

The combination of these considerations aptly illustrates the archival challenges that result from the core characteristics of the new, web-native objects of scholarly communication. It also illustrates that the atomic perspective that underlies journal archiving is inappropriate for archiving in the era of the web of objects. Journals can be archived one by one, independent of each other. The fixed nature of their content and of their references guarantees that each article's information context can be recreated by visiting journal archives. The web of objects calls for another archival paradigm that inherently takes the interlinked and dynamic nature of the new scholarly objects into account.

Web archiving can serve as inspiration with this regard, especially since all objects of scholarly communication reside on the web and link to other web resources, both traditional articles and novel objects. When archiving web pages, web archives will not just archive a page's HTML but also embedded resources such as images and linked resources. As such, the information bundles that web archives collect are not dissimilar from the interlinked compound scholarly objects. And, web archives allow revisiting pages as well as their linked context as they existed at some time in the past. The Memento protocol [4] even supports including multiple web archives as well as versioning management systems in the recreation of the past. This is not dissimilar from the need to revisit a scholarly object and its linked context (see Figure 3).

Although the web archiving paradigm seems appropriate for the task of archiving the web of objects, the current practice is not sufficient to achieve accurate recreations of dynamic interlinked objects. This is aptly illustrated by Figure 5 drawn from a paper that explores temporal incoherence of pages in web archives [1]. The figure shows a page recreated by the Internet Archive. Although the page is the weather report for the city of Varina in Iowa on October 9th 2004, it doesn't take too much imagination to find similarities with a scholarly object, for example, by its inclusion of graphs, data points, data visualizations. The figure critically reveals that the page has been recreated by means of archived web resources with archival dates that range between 20 days prior and 9 months after October 9th 2004, a result of web crawling strategies and, likely, archive de-duplication processes. The recreated page is temporally incoherent and actually never existed in the way the web archive recreates it.

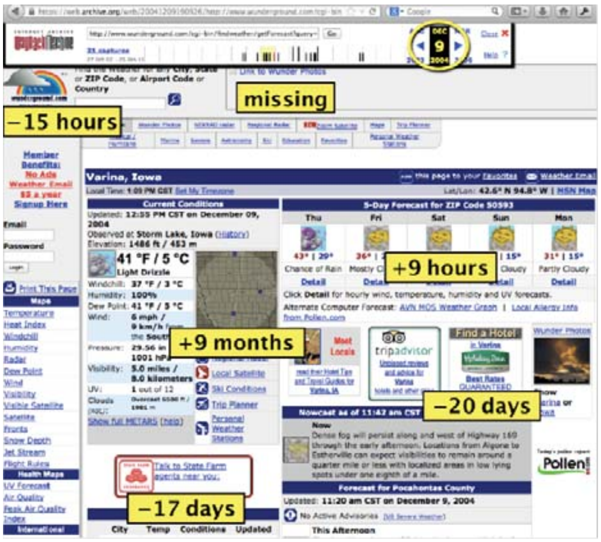


Figure 3: Temporal incoherence of an archived web page

On-demand web archives such as archive.is[18] do not exhibit this deficiency as they collect a resource and its embedded resources at the very moment a user requests it. Although this type of web archive typically does not collect linked web pages, the snapshot approach is more aligned with the requirements for archiving the web of objects. Still, the suggested trend towards content-driven archiving means that constituent or linked resources of a scholarly object can not be archived in a single place, but rather in specialized, distributed archives. This requires some sort of orchestration driven, among others, by content type of the archival process. As is the case with regular web archives, the Memento protocol could serve as the interoperable glue to recreate specific states of objects from snapshots available in multiple archives.

5. THE FUTURE OF ARCHIVING

As described, the emerging web of objects has fundamentally different characteristics than its predecessor, the system of journals. Several of those characteristics, especially the interlinked, dynamic, and heterogeneous nature of the objects suggest the need for a different archiving paradigm. The web archiving paradigm can provide inspiration as it is based on the understanding that, on the web, resources are interlinked and their interpretation critically depends on their network context.

5.1 Infrastructure considerations

Figure 4 provides a high-level view of a future scholarly infrastructure based on the above discussion, and inspired by [5]. Contributing to the planning and building of such an infrastructure is the subject of ongoing work by the authors and their colleagues.

A researcher conducts some of her research on private infrastructure, personal computing facilities. Since the objects created in this environment reside in local namespaces, their inclusion in a web-wide archival solution is hindered. As a result,

[17] http://hiberlink.org/

[18] http://archive.is

from a scholarly system perspective, objects in private infrastructures are ephemeral, and cannot be considered parts of the scholarly record.

A variety of incentives lead the researcher to move her objects from the private infrastructure to the web-based recording infrastructure. These include sharing with self across computing platforms, sharing with a team of collaborators, or even complying with a requirement from a funding agency. As described, recording platforms have attractive features when it comes to fulfilling the registration, certification and awareness functions of a scholarly communication system, but archival platforms they are not. However, because the recording platforms are embedded in the web, objects now reside in a global namespace and are network accessible. Hence, they are within reach of web-scale processes aimed at selectively moving objects from the recording infrastructure into the archival infrastructure, and hence into the permanent scholarly record.

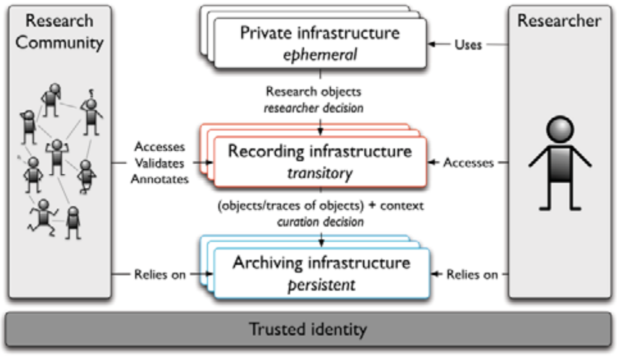


Figure 4: High-level view of a future scholarly infrastructure

Core aspects of these processes include the ability to snapshot the state of interlinked objects at specific moments in their lifecycle, to transfer these snapshots from a variety of recording platforms to appropriate distributed, content-driven archives, and curatorial policies aimed at deciding what should be archived when.

Underpinning the entire infrastructure is a trust component that provides assurances regarding identity and authorizations.

5.2 Curatorial considerations

Assuming the existence of web-scale processes that are able to transfer objects from their operational state in the recording infrastructure to an archival state in the archiving infrastructure, significant questions of a curatorial nature remain. Indeed, in order for the archival infrastructure to stand a chance at sustainability, significant curatorial filters will be required.

A first consideration pertains to what the archival object should be, or, to use the above terminology, what the nature of a snapshot is. For certain objects this may be a copy of the actual object, for others metadata that describes the state, or provenance information that can be used to recreate the state.

A second consideration pertains to the inputs that trigger the transition from operational to archival state. A variety of options present themselves with this regard. In the conservative scenario of Figure 4, the submission of a manuscript or the publication of a paper may launch a process aimed at collecting snapshots of all linked resources. Network-derived metrics, such as altmetrics[19] that measure impact of scholarly objects by means of their presence in the social network flow, or their use, could be used to guide a decision. Decisions could be on-demand, initiated by a researcher as a means to preserve what she considers an important state of one of her own objects, or to safeguard the state of a colleague's object before starting to build on it. It might also be worthwhile to introduce a level of randomness in the decision making to increase the chances of capturing objects that might be serendipitously interesting in the future.

A third consideration is around how the archiving decision is made. Given the vast number of objects that will reside in the recording infrastructure, largely automated decision making driven by heuristics like the aforementioned ones seems essential.

The implications of these considerations are being worked through with archival specialists at DANS[20].

6. CONCLUSIONS

This paper has provided a perspective of a future scholarly communication system, called the web of objects, and has focused on the impact of that system on the fulfilment of the archival function. A core observation was the increased use of web-based recording platforms that excel at registration, certification, and awareness but provide no guarantees regarding archiving. Hence, the introduction of web-scale processes aimed at transferring objects from recording platforms to appropriate archives, subject to curatorial filters was proposed.

Archival infrastructure that underpins research communication needs to be trustable and hence sustainable for the long term. Sustainability, in light of the heterogeneity and number of objects requires a distributed approach. A distributed archival approach to present the web-based scholarly record in a uniform, interconnected manner, requires interoperability and thus standards.

7. ACKNOWLEDGMENTS

The authors would like to acknowledge DANS for bringing them together to work on the ideas presented here.

8. REFERENCES

[1] Ainsworth, S., Nelson, M. L., and Van de Sompel, H. 2014. A Framework for Evaluation of Composite Memento Temporal Coherence. arXiv:1402:0928

[2] Roosendaal, H. E. and Geurts, P. A. Th. M 1997. Forces and functions in scientific communication: an analysis of their interplay. In *Proceedings of CRISP 97*. http://www.physik.uni-oldenburg.de/conferences/crisp97/roosendaal.html

[3] Van de Sompel, H. et al., "Rethinking Scholarly Communication: Building the System that Scholars Deserve", *DLib* (September, 2004). doi:10.1045/september2004-vandesompel

[4] Van de Sompel, H., Nelson, M. L., Sanderson, R. 2013. RFC7089 HTTP Framework for time-based access to resource-states – Memento. http://tools.ietf.org/rfc/rfc7089.txt

[5] Treloar, A. and Harboe-Ree, C. (2008). "Data management and the curation continuum: how the Monash experience is informing repository relationships". *Proceedings of VALA 2008*, Melbourne, (February 2008). http://www.valaconf.org.au/vala2008/papers2008/111_Treloar_Final.pdf

[19] http://altmetrics.org/

[20] http://dans.knaw.nl/

Building Information Modeling – A Game Changer for Interoperability and a Chance for Digital Preservation of Architectural Data?

Michelle Lindlar
German National Library of Science
and Technology (TIB)
Welfengarten 1B
30167 Hannover
+49 511 762 19826
michelle.lindlar@tib.uni-hannover.de

ABSTRACT

Digital data associated with the architectural design-and-construction process is an essential resource alongside -and even past- the lifecycle of the construction object it describes. Despite this, digital architectural data remains to be largely neglected in digital preservation research – and vice versa, digital preservation is so far neglected in the design-and-construction process. In the last 5 years, Building Information Modeling (BIM) has seen a growing adoption in the architecture and construction domains, marking a large step towards much needed interoperability. The open standard IFC (Industry Foundation Classes) is one way in which data is exchanged in BIM processes. This paper presents a first digital preservation based look at BIM processes, highlighting the history and adoption of the methods as well as the open file format standard IFC (Industry Foundation Classes) as one way to store and preserve BIM data.

General Terms

Communities, preservation strategies and workflows, specialist content types

Keywords

Architectural 3D data, Building Information Modeling, 3D preservation, IFC

1. INTRODUCTION

Mankind's desire to construct buildings – and with that the history of architecture – can be traced back to the Neolithic period. Buildings do not only provide shelter, but serve many functions in our life. Cities may be easily identified through a characteristic building, such as the Eiffel Tower or the Sydney Opera House. Naturally, design and construction of buildings remains one of the largest sectors in the 21st century – in the US alone, the annual spending on construction in 2013 was at $898.4 billion [1].

The construction of "standard" objects, such as residential buildings or smaller to mid-size non-residential structures, are as much a part of the design-to-construction process as projects which focus on the combination of aesthetic expression, physical principles and innovation, such as in the case of the "3D print canal house", a research- and building site in Amsterdam where architects are for the first time testing the use of 3D printed building parts in design and construction.[1] Another area are large (total cost more than $10 million) and mega-projects (total cost over $1 billion), such as the new Istanbul airport with a planned capacity of 150 million passengers per year [2].

Architectural records may be archived for different purposes and reasons, three of which should be mentioned here: The first case is that of regulatory requirements, which require the deposit of design and construction records, especially in the case of publically funded buildings, to a regional or national body such as a national archive. The second case is that of the building owner or facility manger, who relies on the availability of the information for reconstruction or simple maintenance purposes. The last example is that of the architectural records being preserved by a library, archive or museum for the historic value or the significance of the construction object or the architect. Prominent examples of special collection libraries for architectural content include the Avery Architectural & Fine Arts Library at Columbia University[2] or the RIBA Library of the Royal Institute of British Architects[3].

This paper gives an insight into the various stages at which architectural data is produced and used along the building's lifecycle. The lifecycle view provides an understanding of the different actors which function as producers and consumers – and therefore also as the designated community for the digital data produced. Until recently, the domain has been dominated by a lack of interoperability which has lead to a decline of productivity. While the concept of Building Information Modelling (BIM) has existed for over 30 years, it has only been adopted recently. A brief history of the process and its adoption shall give a better understanding of the idea behind Building Information Modelling. Lastly, IFC (Industry Foundation Classes) is introduced as an open and standardized way to exchange Building Information Modelling. A description of the format against sustainability factors and a brief risk assessment puts the

[1] http://3dprintcanalhouse.com/

[2] http://library.columbia.edu/locations/avery.html

[3] http://www.architecture.com/LibraryDrawingsAndPhotographs/Home.aspx#.U0g9tqJcLQE

content type and format further into a digital preservation perspective.

2. RELATED WORK

Despite the economic value of digital architectural data and despite the significance that digital architectural records may hold in the cultural heritage context, very little research has been conducted regarding the digital preservation of the material. First efforts in this direction were made by MIT's Façade (Future-proofing Architectural Computer –Aided Design) project[4], which ran from 2006 – 2009. The project focused on proprietary CAD (computer aided design) files which were deposited in the institution's DSpace based preservation repository. The project pointed out the heterogeneous software landscape in architectural practice and the legal restrictions connected to the proprietary formats as two of the biggest problems in the preservation process. Façade reached the conclusion that the best preservation strategy would be to preserve 4 versions of the object: (1) the original submitted digital object, (2) an access copy, in particular 3D PDF (3) a full "preservable standard format", in particular STEP or IFC[5] (Industry Foundation Classes) (4) a "preservable standard format" containing just the geometry, in particular IGES [3].

While not dealing exclusively with architectural CAD data, the 2013 DPC (Digital Preservation Coalition) report "Preserving Computer-Aided Design" comes to a similar result, suggesting that archives should keep the original CAD file and migrate to at least one vendor-neutral format, where in particular STEP standard based formats are pointed out as being suitable [4].

Both points of reference – MIT Façade as well as the DPC report – focus on the CAD object as the preservation starting point, therefore following an object-centric as opposed to a process-centric approach. A process-centric approach helps us to understand different players involved in production or usage scenarios of the data. This will eventually lead to BIM – a practice that had not been as widely adopted during the running time of the MIT Façade project as it is today.

3. A LIFECYCLE VIEW

A lifecycle view of a typical building is a helpful tool to understand the various stages at which architectural data is created and used. The beginning of the lifecycle is marked by the conception of the structure to be built, while the demolition or the re-purposing mark the end or re-start of the cycle. The steps in between may be broken down into two high-level categories: construction and use. These high-level categories signify the temporal aspect of the lifecycle in regards to data production and data re-use – while the construction phase, which is simultaneously the part of the cycle where the most data about the building is produced, lasts on average about 2.5 years, the usage phase, where data from the construction phase is re-used, lasts about 60 or more years.

A more granular look at the two main stages sheds light upon the different actors involved in the construction and usage processes. The concept and the design phases are typically led by the architect who designs the building. Based on this initial design, further actors are involved in the pre-construction phase to define specific needs to various aspects of the building, such as structural engineers or HVAC (heating, ventilation, air conditioning) engineers. Furthermore, information regarding cost or projecting details such as time schedules are defined in preparation of the construction process. During the construction phase, part of the data produced so far is used by the site or construction manager to organize and monitor the physical construction phase itself. At this stage, the construction management as well as the construction companies produce further data, which documents the as-build state. Besides project management information and costs, this may include specific product information or further specifications of the original design. During the hand-over stage the produced data can serve as a verification measure of the construction vs. design process – moreover, it forms the necessary documentary basis for the operation of the building. For large objects such as, e.g., hospitals, hotels or large-scale office buildings, facility management companies rely on complete and exact data regarding various building parts to ensure economically efficient and safe operation and maintenance of the structure. During the use-phase new data may be created for various reasons, such as in the case of producing documentation for regulatory decrees, e.g., in form of required documentation for new fire safety regulations, or in the case of the documentation of minor modifications, such as the installation of a new parts within the heating system or the tearing down of a non-bearing wall to create a larger room.

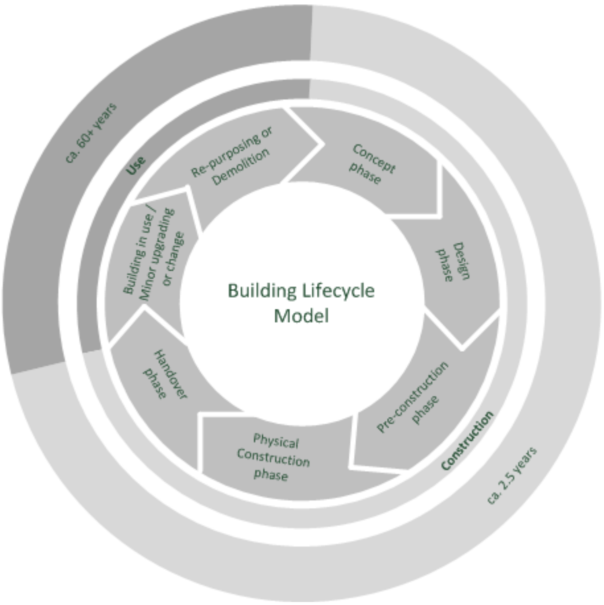


Figure 1. Building Lifecycle

As suggested above, the vast majority of data about a building is produced during the "construction" stages. Table 1 shows the amount of paper-based information that typically occurs for large-scale projects (construction cost exceeding $ 10 Million). In addition to showing the amount of documentation produced, the table displays the fragmented nature of the construction domain resulting in the comparatively high number of different companies involved in the process. Based on EU industry sector

[4] http://facade.mit.edu

[5] The project did not look at STEP and IFC in the context of BIM data, as the process had only just begun to establish itself in the architectural practice [3].

statistics, companies with fewer than ten employees accounted for 90% of the European construction industry workforce in 2005 [5].

Table 1. Typical numbers for large-scale projects with a cost of $10 Million [6]

Number of pages in documents	56,000
Number of individual participants involved	850
Number of companies involved (including suppliers and sub-sub-contractors)	420
Number of types of documents generated	50
Number of banker boxes to hold project documents	25

How can cooperation between and seamless integration of so many actors be realized in a business process as diverse as the design-to-process one? That the situation is not ideal has been displayed in various ways – one being a 2004 analysis conducted by Teicholz [7], where the 1964 to 2004 productivity index of the construction domain was compared against that of all other non-farm labor domains. While productivity for the non-farm labor domains had gone up steadily, that of the construction domain had actually decreased. In other words: construction projects of 2004 cost significantly more hours per dollar than they did in 1964. Teicholz sees one of the main reasons for the productivity decline in the nature of ICT stand-alone-system developments of the various actors involved in the design and construction process. While each sub-domain may use state-of-the-art systems in their own right, there is a lack of interoperability which in the worst case leads to information being exported from a digital system to paper documents and then manually re-imported from there [7]. A 2013 analysis of the UK's construction industry's supply chain suggests that the situation has not improved since Teicholz observations made in 2004. The 2013 analysis shows poor quality information and incompleteness of design as a major cost factor, in some reported cases being as high as 25% of the overall building cost [8].

The lack of cooperation in the digital age is a much reported issue in architectural and construction related research [7], [10], [11], [12]. Hitchcock and Wong, to give one example, point out that in the case of energy simulation building models, the lack of robust data exchange methods has lead to a practise, where data is collected from various sources and transformed based on professional expertise and a rules-of-thumb approach instead of a standardized one. This often leads to a range of different possible energy simulation building models for the same initial object [9].

The fragmented nature of the documentation of the architectural design-to-construction records naturally poses a challenge not only for the cross-sectional usability, but also for the preservation process of the digital information associated with an architectural design- and construction project. As described above, actors involved in the design-to-construction process may use their domain-specific and often proprietary monolithic software solutions to produce information.

4. BUILDING INFORMATION MODELLING (BIM)

A solution to the lack of interoperability, to incomplete data and to the low productivity associated with these problems is seen in a widespread adoption of Building Information Modeling (BIM) as a consequent model throughout a building's lifecycle [13].
While the acronym BIM is most frequently translated as "Building Information Modeling", it may be resolved in the following ways: [10]:

1. as Building Information Modeling, which describes the business process of generating and maintaining semantically rich digital objects which contain geometry and layout as well as information on material, cost estimation and scheduling.
2. as Building Information Models, the instantiation produced by the process described in (1)
3. as Building Information Management, which refers to the organization and control of the processes associated with processes in (1), the digital objects in (2) and their utilization along a building's lifecycle

A good definition of the term is given by Nederveen et al. [14]:
*"a **Building Information Model** is an information model of a building (or building project) that comprises complete and sufficient information to support all lifecycle processes, and which can be interpreted directly by computer applications. It comprises information about the building itself as well as its components, and comprises information about properties such as function, shape, material and processes for the building life cycle".*

4.1 Brief History of BIM

The idea behind BIM dates back to the 1970s and 1980s. Early terminology used to describe the concept differed. Charles Eastman first proposed the idea behind what is today known as BIM in 1975, describing a prototype of a "Building Description System" which aimed to combine the advantages of manual drawings and physical models in a computer graphics based system. The "Building Description System" recognized a number of facts which formed the foundation of what is today known as BIM, such as the fact that every element of a building essentially consists of three types of descriptions – (1) shape (2) location and (3) a list of properties – and that every element may occur several times in a building, differing in only the location descriptor [16].

From there, research and development in the USA and Europe further developed the idea while assigning different terminology to the concept. While the term "Building Product Model" established itself in the USA, in Europe, the term "Product Information Model" was used. Robert Aish specified the concept further in 1986, including most of the cornerstones that today make up BIM and giving it the label of "Building Modelling" [13].

The full term "Building Information Modelling" was introduced in 1992 by G.A. van Nederveen and F. Tolman, who focused on the modelling of different views of a building in order to support various stakeholders' needs [17].

Despite the fact that the concepts of BIM had been represented in AEC software as early as 1987[6], the terminus coined by van Nederveen and Tolman remained dormant for 20 more years until a 2002 Autodesk Building Industry Solution White Paper entitled "Building Information Modeling". Autodesk described Building Information Modelling as it's "strategy for the application of information technology to the building industry" [18]. At the core of Autodesk's strategy was the inclusion of digital databases, which shall facilitate collaboration, better change management, as well as easier reuse of information.

In the context of digital preservation it is interesting to note that the white paper states two preservation cases:

1. The system shall "capture and preserve information for reuse by additional industry specific applications"

2. The system shall capture audit trail information about changes made by all team members and preserve it "for as long as this information is useful" [18]

The fact that the terminology BIM was then picked up by the two other large software companies on the AEC design market – namely Bentley Systems and Graphisoft – can be attributed to industry analyst Jerry Laiserin. Laiserin suggested a global adoption of the term "Building Information Modelling" reasoning that "CAD is no longer sufficiently descriptive of the breadth and depth of the design process" [19] and he gave the CEOs of the respective companies a forum to exchange opinions on the adoption of the term [20],[21],[22].

4.2 Moving beyond CAD – key features

As previously mentioned, building models describe a building as a structured set of intelligent components which in themselves are characterized on three levels: they are associated with a computable graphic / are spatial, they are described through data attributes and they may be modified through parametric rules. The data which describes the elements shall be consistent, non-redundant and include behavior, such as information needed for energy simulations [13].

As opposed to other industries' application of parametric based modeling, BIM software comes with a pre-defined set of building elements, which are broken down into smaller categories or "families" at which level they may be modified or extended by the user. These families are described in parametric relationships to each other, enabling the software to coordinate and manage the changes made to the building model. To give an example: a floor is attached to a wall – if the floor size is changed, the wall moves accordingly. These conditions are defined in rules – to again pick the example of a wall: rules include checking that doors and window locations lie completely within a wall and that the locations of doors and windows do not overlap each other.

Building Information Modeling allows the generation of different views – or representations – based on a single building model, e.g. in form of a 2D or a 3D representation or in form of a design view and a view of the HVAC (heating, ventilation, air conditioning) parts. Figure 2 shows an example of different views generated from the same model.

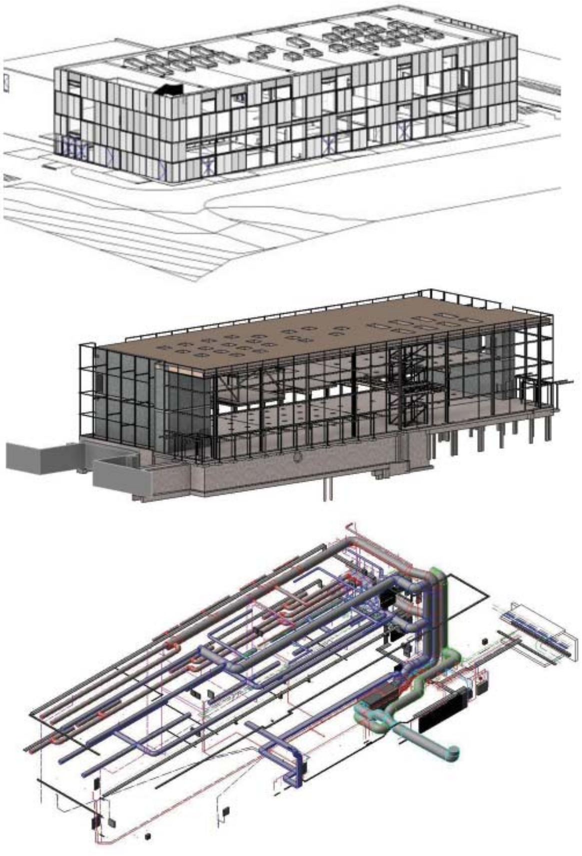


Figure 2. Different views of the same object – architect (top), construction engineer (middle), HVAC engineer (bottom)

The different views shall allow the different actors to remain an easy access to the file on a level that feels "familiar" to their domain. This interoperability enabler shall lead to accurate and complete data, thus supporting the design-to-construction process down to the handover phase. In the usage phase of the building's lifecycle, complete and detailed data shall greatly benefit facility management in efficient and sustainable operation of the building [13].

4.3 BIM Adoption

While Nederveen et al. specifically included a building's life-cycle-long support in their earlier quoted BIM definition, they also pointed out in 2010 that the definition may be considered as a future outlook which is currently far from common practice [15]. But what does the situation look like today?

While the popularity of the search term "Building Information Modeling" suggests a growing interest in the subject matter (see Figure 3), table 2 shows that BIM is seeing growing adoption and is as of today a required process for publically funded construction projects in a number of countries.

[6] In a 2003 issue of the LaiserinLetter, Graphisoft's then Vice President for Architecture Chris Barron described Graphicsoft's adoption of the concepts of BIM in ArchiCAD's "Virtual Building" approach, which dates back to 1987. [22]

Table 2: BIM and IFC adoption

Country	BIM Status	IFC Status	Driver
Australia	Not mandatory	Not mandatory	Association driven Driven by public organisations like the Australian Construction Industry Forum; successful BIM implementations for maintenance of some large objects like the Sydney Opera House.
Denmark	Mandatory (partially)	Mandatory (partially)	Government driven Regulations starting April 2013 were passed by the Danish Building and Property Agency[7] and are required for construction projects which are at least 50% state financed, exceed overall construction cost of 5 Million DKK or are results of architectural competitions. BIM and IFC are both mandatory for those objects. Triggered by the 2007 government initative "Det Digitale Byggeri" (Digital Construction) some Danish government / state level agencies had previously already been requiring BIM, and specifically IFC.
Finland	Mandatory	Mandatory	Government driven Both BIM and the delivery in the IFC file format are mandatory for government projects since 2007 as per Senate Properties[8] regulations.
Germany	Not mandatory	Not mandatory	Association driven A first government initiative was the recently published "BIM recommendations for Germany", intiated by the Federal Institute for Research on Buidling, Urban Affairs and Spatial Development[9].
Hong Kong	Mandatory	Not mandatory	Government driven BIM will be mandatory for all Hong Kong Housing Authority[10] projects from 2015 (for some, from 2014) on. While the inclusion of open standards is encouraged, no specific requirements in regards to IFC are made.
Netherlands	Mandatory (partially)	Mandatory (partially)	Government driven Rijksgebouwendienst (Rgd)[11] of the Dutch Ministry of the Interior has been requiring BIM for only some of the publically funded projects since 2012. For those projects where BIM is required, BIM extracts including the IFC model alongside CAD drawings and measurement data, calculations, etc. are expected per the Rgd BIM Standard of 2012.
Norway	Mandatory	Mandatory	Government driven The government organization "Statsbygg"[12] has been requiring BIM as well as IFC for all government construction projects since 2010.
Singapore	Mandatory (partially)	Mandatory (partially)	Government driven The Building and Construction Authority (BCA) has passed regulations requiring BIM for new building projects exceeding 5,000 sqm in size. The BCA developed e-submission system for BIM requirements "CORENET" implements the IFC model.[13]
United Kingdom	Mandatory	Mandatory	Government driven The Government initiative "Government Construction Strategy"[14] requires BIM for all government construction projects from 2016 on. Models will need to be available in the COBie UK 2012 schema[15], which may be derived from an IFC MVD.
USA	Mandatory	Mandatory	Government driven General Service Administration (GSA)[16] regulations have been requiring BIM for government construction projects since 2008. For those projects, the availability of the native CAD format and the IFC object are required. The Army Corps of Engineers is a second government body which made BIM mandatory for all projects

[7] http://www.bygst.dk

[8] http://www.senaatti.fi/en

[9] http://www.bbsr.bund.de/

[10] http://www.housingauthority.gov.hk/en/index.html?url=/en/

[11] http://www.rgd.nl/english/

[12] http://www.statsbygg.no/System/Topp-menyvalg/English/

[13] https://www.corenet.gov.sg/

[14] https://www.gov.uk/government/publications/government-construction-strategy

[15] http://www.bimtaskgroup.org/cobie-uk-2012/

[16] http://gsa.gov/bim

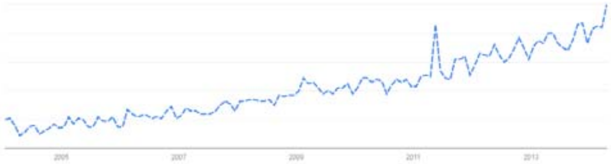


Figure 3: World-wide Google search term development for "Building Information Modeling" 2004-2014[17]

It has also been recognized, that the use of BIM technology and processes significantly changes the relationships, communication and collaboration ways of the actors being involved in the design-to-construction process [13]. This is closely tied to the third acronym interpretation of BIM given in the introduction to this chapter: Building Information Management.

On an organizational level, the role of the "BIM Manager" is being given more attention, with government based guidelines – such as Hong Kong's roadmap for a strategic implementation of BIM – specifically suggesting a BIM manager in every project "to develop integration mindset and whole lifecycle systems' mindset to project participants" [25].

On an ICT level, the integration and collaboration need is being met through model servers, which manage file exchange between the different actors as well as versioning and consistency. These model servers allow the import from and export to CAD & BIM desktop tools and may furthermore integrate product databases provided by vendors or large agglomerated databases like the nbs (National Building Specification) National BIM Library[18]. Most model servers will store the information on the models in databases, which are used to generate the views for the specific needs pertaining to the respective actor – such as a view for the structural engineer as opposed to the facility manager (see figure 3). The respective actors work with the models in their own sub-domain specific software and upload the results to the BIM Server, where the information on the construction object is then synchronized.

BIM integration in software can be divided into two approaches: One is a vendor based solution, where a vendor will support BIM integration through different software solutions within a suite. An example for this is Autodesk's BIM solutions, where models can easily be exchanged between different available software modules for architectural design, construction and facility management.[19] This vendor-based BIM process is sometimes referred to as "closed BIM". While it comes at the price of complete dependency on the software vendor, it allows tight integration and the full exploitation of features that single software systems of a suite entail. In Figure 4, the given examples for closed BIM data exchange include the native BIM formats DWG (AutoCAD Drawing), RVT (Autodesk Revit Project File), DGN (MicroStation DesiGN File) and GSM (Graphisoft ArchiCAD File).

The second approach facilities collaboration between the different involved actors through the use of publically available standards as exchange methods between different software platforms. This method is sometimes referred to as "open BIM". While this approach comes at the price of most likely not being able to maintain some of the functionality that the source software included for the original file format, it allows for a much higher degree of flexibility between the different actors without any software vendor dependency.

A few of the exchange formats shown in figure 4 are proprietary exchange formats, of which DXF (Data eXchange Format) is the most common one. DXF is a format defined by Autodesk which has become somewhat of the smallest common denominator in the exchange of vector data between CAD systems. The problem with the DXF format is that it typically changes with every new release of the AutoCAD family [23].

A second group of file formats shown in figure 4 can be classified as access formats, as they are stable and openly available formats which are supported by a number of readily available viewers while only exposing a fraction of the BIM information (e.g., JPEG, PDF, PDF 3D, OBJ).

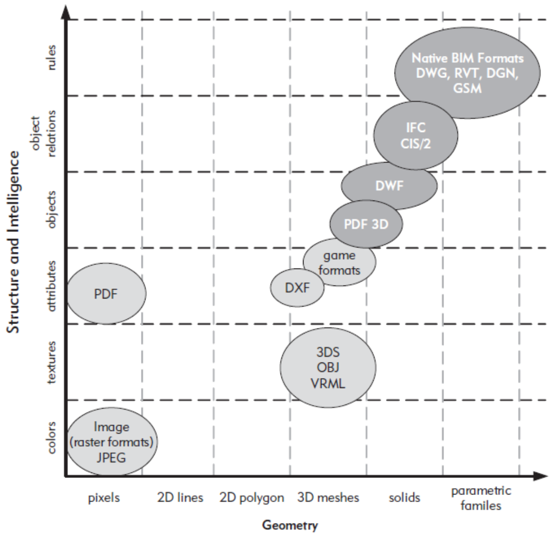


Figure 4. Level of geometry, structure and intelligence in potential data exchange formats [13]

Currently only two open exchange formats exist which fully support BIM: IFC (Industry Foundation Classes) and CIS/2 (CimSteel Integration Standard). Both standards are based on ISO-STEP technology (see chapter 5.1), are human and machine readable, are standardized, publically recognized and widely used. While CIS/2 supports structural steel design only, IFC is targeted at the entire BIM spectrum. A mapping between the two standards has been developed to allow for interoperability.[20] Widely used

[17] retrieved April 9th 2014

[18] http://www.nationalbimlibrary.com/

[19] http://www.autodesk.com/solutions/building-information-modeling/overview

[20] The mapping is available at website of the National Institute of Standards and Technology: http://www.nist.gov/manuscript-publication-search.cfm?pub_id=861673

"Open BIM" model servers, such as the "BIMserver"[21], use IFC as the data exchange format.

While XML (Extensible Markup Language) is a file format often used for interoperability reasons and data exchange, it currently only finds usage for smaller sections of the BIM process. An example for this is gbXML (Green Building XML) which is a schema supporting the data contained in BIMs to engineering analysis tools. Semantically, gbXML could be considered a subset of IFC, as it does not contain relevant information which cannot be modelled in IFC [24].

A general problem that pertains to any exchange format is the fact that it relies on stable import and export mechanisms into and out of often proprietary source systems. These mechanisms need to be checked consistently after updates of the source software as well as after updates to the exchange format.

5. INDUSTRY FOUNDATION CLASSES (IFC)

While the term "Building Information Modelling" was not widely adopted until 2002, as described in the previous chapter, the strife for interoperability in the AEC (architecture, engineering, construction) / FM (facility management) domains is much older. The need for easily exchangeable and reliable data has put forth the development of the Industry Foundation Classes (IFC) standard.

IFC can be described as a hierarchical object sub-typing structure, in which objects are nested in an entity tree and each entity is described with attributes. The attributes may describe an object's material, behaviour (e.g., thermal characteristics) or contextual properties (e.g., weather data) as well as process related characteristics such as time, fire safety regulations, building use or projected cost [13].

The latest IFC version (IFC4) contains 766 entities, meaning that 766 different concepts or objects exist in the schema, each of which can be instantiated numerous times within a model, be described with attributes and be set in relation to other entity instances [26].

As of today, the IFC data model is the only comprehensive, public, non-proprietary and well-developed data model which supports the full design-to-construction process [13].

5.1 Brief History of IFC

The "standard behind the standard", so to speak, is STEP, which is standardized as ISO10303. The idea behind the STEP standard itself dates back to 1984 when the decision to develop an open product modeling standard which could serve the needs of a wide variety of industrial and manufacturing industries was made by the ISO TC184/SC subcommittee. This was to be achieved by central core elements, which domain specific application protocols could be built upon, thus avoiding redundant standard development across several domains and paving the way for easier collaboration between different industrial manufacturing industries. At the heart of the common core of STEP was the idea of a robust data model describing concepts like relationships, attributes, constraints and inheritance [12].

The method to describe these concepts was realized in form of the EXPRESS information modeling language, which functions as the core of various other STEP data models, for example the aforementioned CIS/2 or for application protocols of other domains, for example LOTAR[22] for the aerospace and defense industries. File formats and schemas based on STEP need to be based on a machine readable modeling language instead of a binary file format. The language should include clear data declarations but also include rules and constraints to model procedural requirements. The standard requires the mapping to be applicable to different implementations, namely a text file format ("Part-21"), a SQL and object based database implementations as well as an XML schema ("Part-28"). Lastly, it should allow for the development and inclusion of sub-models to support the needs of specific domains [13].

While the initiation of STEP development dates back to 1984, the first STEP standard was not released until 1994. For the AEC/FM industry this was too slow-moving and unresponsive to the domains' needs which lead them to undertake their own efforts in driving interoperability through format development and standardization. It may seem surprising that the development of IFC was at it's base a process driven by software companies. Under the lead of Autodesk, 12 U.S. based industry and software companies founded the IAI (Industry Alliance for Interoperability) in 1994 with the aim to drive tool and standard development supporting the data exchange amongst actors involved in planning, construction and maintenance of a building. In 2005 the IAI changed its name to buildingSMART[23] [12].

The years 1994-1999 can be considered the early days of IFC prototyping. Format version 1.0 focused solemnly on the architectural part of the building, while IFC version 1.5.1. was the actual first implementation in a BIM software. While the efforts so far had been mainly conducted in the U.S., IFC version 2.0 was the first true international prototype, incorporating work of newly established international IAI charters. IFC2.0 incorporated schemas for cost estimation, building services and construction planning and can be considered the last prototype of the IFC format development [12]. The file format versions 1.0 to 2.0 are now considered obsolete and are no longer supported [26].

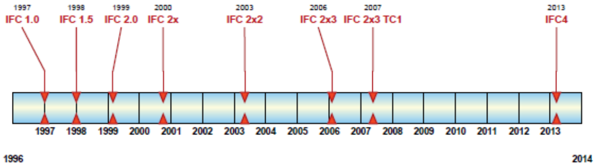


Figure 5. Timeline overview of IFC format releases [27]

The first stable "production" release was IFC2x, released in 2000. The next major release IFC2x2 in 2003 added new domain areas, while IFC2x3 in 2006 addressed mainly quality issues of the model. Even though STEP ISO10303 conformity is still fulfilled, IFC became its own ISO standard in 2013: ISO16739. The same

[21] http://www.bimserver.org/

[22] http://www.lotar-international.org/

[23] http://www.buildingsmart.org

year, the most recent version was released: IFC4. The IFC4 release enables new BIM workflows which have been developed within the domain since the 2x developments, including GIS interoperability and enhanced thermal simulations. Furthermore, ifcXML schema description, which was previously conducted in parallel to the text file format IFC-SPF, is now included in the general version specification. Simultaneously, the XML Version has been improved significantly, reducing the needed lines down to 14% of what it was at in IFC2x3 XML, making it 6 times more efficient [27].

5.2 IFC Adoption

As mentioned before, IFC is today a widely accepted standard [13], [12]. Seven out of the eight national regulatory bodies which require BIM and are documented in table 2, also require the documentation of the design-to-build process using the IFC file format standard. The only exception to this is Hong Kong, who is just now in the process of realizing BIM regulatory requirements and mentions the focus on open standards, however, without implicitly pointing towards IFC. It will remain to be seen, whether IFC will be picked up in the requirements there as well [25].

On a software level, a number of freely available IFC viewers are available, such as the Solibri Model Viewer[24] or the DDS IfcViewer[25]. Furthermore, the IFC core model is today supported by more than 150 software tools.[26] To make the stability of import and export routines into and out of CAD or other systems transparent, the buildngSMART foundation maintains a certification process for third party applications. Here, software developers may certify their application towards the support of an IFC version. Currently, certification is available towards the IFC2x3 standard and has been started or completed for 31 different applications.[27]

6. DISCUSSION

As demonstrated in Figure 4, the IFC file format preserves a high degree of the BIM object's intelligence and geometry. While some parametric information as well as rule functionality of the source systems may be lost, a growing adoption of the file format has built a community which addresses these questions in processes such as certification procedures for import and export routines out of monolithic domain-specific software. Furthermore, the file format is supported by a growing number of open source tools for file analysis, viewing and manipulation.[28]

BIM certainly simplifies the process of capturing a building's documentation by containing a lot of information which was previously only available in a spread-out manner across numerous pieces of documentary evidence. It furthermore fulfills a lot of the needs of the various designated communities aligned around the building's lifecycle. IFC seems to further support this process from a preservation view by doing so in an open, standardized and well adopted way.

Measuring the file format against well recognized sustainability factors will give further insight into digital preservation suitability of the file format [28].

6.1 File Format Sustainability

The sustainability factors described here are based on an analysis conducted as part of the DURAARK (Durable Architectural Knowledge) project [28]. It needs to be noted that three representation forms are available for the IFC file: in addition to the previously mentioned clear-text renditions IFC-SPF (IFC STEP Physical File, .ifc) an IFC XML (.ifcxml) and IFC-ZIP (.ifczip) version is available, which compresses either IFC-SPF or IFC-XML using PKzip 2.04g compression. The sustainability factors only describe IFC-SPF and IFC-XML, particularly in version IFC4, as the xml specification is included in the general IFC4 specification [28], [27].

Disclosure

As the IFC file format is openly available and standardized, all necessary information about the file formats' design and structure is available.[29] The standardization is clearly written and includes a change log comparing the current to the previous schema. While versions IFC1.0 to IFC2.0 were non-productive prototypes, the version family IFC2x, which was superseded by IFC4 in 2013, remains supported by current tools.

Internal technical characteristics

Following the STEP principles, both the XML and the text based SPF version of the format are human and machine readable, implementation independent and free from encryption. While the schema is certainly complex, this serves the purpose of the nature of BIM. The required different views in the BIM process are supported through the availability of Model View Definitions (MVD), which allow sub-domain views onto the model, e.g. for a structural engineer.

External technical characteristics

As a platform and implementation independent standard, the IFC file format does not depend on specific hardware or software. An IFC file may, however, depend on external information, as product catalogue entries may be referenced through URIs (uniform resource identifier) pointing towards, e.g., a vendor's dataset.

Format Acceptance

IFC is a well adopted standard which is recommended by several national regulatory bodies for the documentation of the design-to-build process for publically funded structures. It is well supported by a large number of tools.

Patent

The IFC standard is open and vendor-neutral; it is free from any patent restrictions.

[24] http://www.solibri.com/products/solibri-model-viewer/

[25] http://www.dds-cad.net/downloads/dds-cad-open-bim-viewer/

[26] http://www.buildingsmart-tech.org/implementation/implementations

[27] http://www.buildingsmart-tech.org/certification/ifc-certification-2.0/ifc2x3-cv-v2.0-certification/participants

[28] http://www.buildingsmart-tech.org/implementation/get-started/ifc-open-source/ifc-open-source-summary

[29] http://www.buildingsmart-tech.org/downloads/ifc

Logical Structure and Transparency

As a clear text format with a well-defined schema, IFC is human and machine readable and transparent to methods for validation of the schema and the file format itself. However, while the schema is rather large to support the entire BIM process, this also requires a certain degree of flexibility with a lot of attributes and entities being optional. This complicates schema validation. Nevertheless, well-formedness on the low-level syntax of the file format itself, which is the main requirement for renderablity and accessibility, is transparent to analysis.

6.2 File Format Risk Assessment

The sustainability analysis put forth two particular problems. The first problem relates to potential problems connected with the validation of the schema. This is closely tied to the flexibility, which is based on a large number of entities, attributes and rules to capture all aspects of the design-to-build process. While validation software for the schema at large exists[30], it checks against the entire schema, which makes it hard for the respective sub-domain actors to find the validation errors that pertain to their scenario. With the release of IFC4 a full integration of the model view definitions (MVD) into the XML structure was announced, which may pave the way for easier view-based validation procedures.

The second problem is that of the digital object's dependency on external resources. This is especially the case when the IFC model is enriched with information from vendor product catalogues or external BIM Libraries and entries are only referenced through a URI. A possible way to address this is to store the respective linked dataset alongside the IFC file. While this would preserve the object in its original state, it would not solve the question of easy traceability of changes in the product database, i.e. if a referenced part such as a door knob is no longer available. This problem is currently being addressed as part of the DURAARK project (Durable Architectural Knowledge), where a semantic digital observatory is proposed, which monitors the external resources regarding their stability and availability and mirror changes into a semantic digital archive [29].

A third problem, which is not a result of the sustainability factor analysis but lies in the nature of file formats which are primarily used as data exchange formats between monolithic systems, is that of the dependency on software vendors to produce accurate import and export routines. In the case of IFC problems have especially been reported in regards to data exchange between different proprietary software [30], [31]. Pazlar and Turk pointed out in 2008 that vendor-side IFC interfaces are not where they should be given the years of development and should not be blindly trusted [31]. Recent efforts in research and development have been targeting this gap through automated metrics for similarity and difference detection [32]. On the user side this means that client side import and export routines in systems have to be checked for every new version of the external software as well as for every new version of the IFC format. BuildingSMART's certification procedure for software vendors is here a good contribution to transparency. Nevertheless, consistent checking of the reliability of the import and export functionalities should be conducted to guarantee completeness of the data. This risk is therefore closely tied to the first risk mentioned – i.e., that of the schema validity as per the different stakeholders – as such validation rules may also assist in the checking of correct data after an export.

7. CONCLUSION

In the introduction, three different purposes for the archival of architectural design-to-construction records were mentioned:

1. Regulatory requirements where objects may be deposited to a regional or national body

2. The building owner or facility manager, who relies on the availability of the information for the maintenance of the object

3. Cultural heritage value of the record based on the structure it documents or on the creator of the object

The lifecycle view of the building itself seen in juxtaposition to the data that is produced and used along the lifecycle showed that in traditional architectural digital practice, where systems were monolithic and data exchange was often conducted in a manual "print-out" way, interoperability – and with that also curation and preservation of the data – posed to be a major problem.

While Building Information Modeling was largely developed and adopted to increase productivity within the design and construction domains, it can certainly be seen as a game changer for digital preservation as well. Table 2 shows a growing number of national bodies which have required BIM to be part of publically financed construction projects. These national bodies tend to stand in close connection to all three of the preservation scenarios mentioned above: as they are national agencies, the data they request will eventually be deposited to a national archive. In the case of the USA this might be The National Archive and Records Administration[31]. Meanwhile regulatory body – such as the General Service Administration - itself is responsible for the maintenance of the building, so the digital object will remain actively used there, most likely within a BIM server which enables the traceability of updates conducted to the building as part of maintenance or minor reconstruction over the course of years. In this context it is very well imaginable that there will be a growing need to implement preservation functionality on top of such BIM servers as the objects' capabilities will be further exploited more and more facility managers and building owners will realize the potential of BIM data availability. Lastly, BIM may ease the preservation of cultural heritage, as the information is available in a central object which significantly eases the maintenance.

While growing adoption of the file format may stand for longevity of the file format and while the standard itself presents strong sustainability factors, this paper has shown that a number of risks do exist. As a growing number of IFC files are already being produced today, the digital preservation and the AEC domains need to engage in joint efforts to identify, understand and manage these risks as early as possible.

[30] A validation tools is included in the buildingSMART certification platform: http://gtds.buildingsmart.com/

[31] http://www.archives.gov/publications/general-info-leaflets/26-cartographic.html#architect

8. ACKNOWLEDGMENTS

This work was partially funded by the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement no. 600908 (DURAARK) 2013-2016.

9. REFERENCES

[1] U.S. Census Bureau News. 2013. *December 2013 Construction at $930.5 Billion Annual Rate*. U.S. Department of Commerce.

[2] Saldiraner, Y. 2012. The new airport in Instabul; expectations and opportunities. *Journal of Case Research in Business & Economics*, 5 .(2012), 1-11.

[3] Smith, M. 2009. *Final Report for the MIT Façade project: October 2006 – Augut 2009*. Massachusets Instituite of Technology.

[4] Ball, A. 2013. *Preserving Computer-Aided Design*. Technology Watch Report. Digital Preservation Coaliton.

[5] ECTP. 2005. *Challenging and Changing Europe's Built Environment*. Technical report. European Construction Technology Platform (ECTP).

[6] Hendrickson, C. 2008. *Project Management for Construction.* Carnegie Mellon University, Pittsburgh, PA

[7] Teicholz, P. 2004. Labor Productivity Declines in the Construction Industry: Causes and Remedies. *AECbytes* Viewpoint #4, April 14th, 2004.

[8] Department for Business Innovation & Skills. 2013. *Supply Chain Analysis into the Construction Industry. BIS Research Paper No. 145.* A Report for the Construction Industrial Strategy, October 2013.

[9] Hitchcock, R. and Wong, J. 2011. Transforming IFC Architectural View BIMS for Energy Simulation: 2011. In: *Proceedings of Building Simulation 2011: 12th Conference of International Building Performance Simulation Association*. (Sydney, 14-16 November 2011). 1089-1095.

[10] Parsanezhad, P. and Tarandi, V. A Holstic Approach to Acquisition of Building Information for a More Effecicient Collaboration. In: *Proceedings of the 7th Nordic Conference on Construction Economics and Organisation 2013*. (Trondheim, 2012). 461-468

[11] Howard, R and Björk, B. 2007. Building information modeling – Experts' views on standardization and industry deployment. *Advanced Engineering Informatics*. 22 (2008), 271-280.

[12] Laakso, M. and Kiviniemi, A. 2012. The IFC Standard – A Review of History, Development and Standardization. *Journal of Information Technology in Construction.* 17(2012), 134-161.

[13] Eastman, C., Teiholz, P., Sacks, R, Liston, K. 2008. *BIM Handbook*. John Wiley & Sons, Inc.

[14] Nederveen, S. van, Beheshti, R., Gielingh, W. 2009. Modelling Concepts of BIM. In J. Underwood and U. Isikdag, *Handbook of Research on Building Information Modeling and Construction Informatics: Concepts and Technologies*. IGI Global

[15] Nederveen, S., Beheshti, S., Willems, R. 2010. Building Information Modelling in the Netherlands; A Status Report. In: *Proceedings of the 18th CIB World Building Congress*, (Salford, United Kingdom, May 2010), 28-40.

[16] Eastman, C. 1975. The use of computers instead of drawings. *AIA Journal*. 63-3 (1975), 46-50.

[17] Nederveen, G., Tolman, F. 1992. Modelling multiple views on buildings. *Automation in Construction*. 1, 3, (December 1992), 215-224.

[18] Autodesk Building Industry Solutions. 2002. *Building Information Modeling*. White Paper. 2002.

[19] Laiserin, J. 2002. Comparing Pommes and Naranjas. *The Laiserin Letter*. Issue No. 15. (December 16, 2002).

[20] Laiserin, J. 2003. Bentley Systems on BIM. *The Laiserin Letter*. Issue No. 18. (January 13, 2003).

[21] Laiserin, J. 2003. Autodesk on BIM. *The Laiserin Letter*. Issue No. 18. (January 13, 2003).

[22] Laiserin, J. 2003. Graphisoft on BIM. *The Laiserin Le*tter. Issue No. 19. (January 20, 2003).

[23] The National Archives. Website Resource on PRONOM PUID 744. http://www.nationalarchives.gov.uk/PRONOM/Format/proFormatSearch.aspx?status=detailReport&id=744

[24] Lam, K., Karaguzel, O., Zhang, R., Zhao, J. 2012. *Identification and Analysis of Interoperability Gaps between Nbims/Open Standards and Building Performance Simulation Tools*. Technicla Report. Carnegie Mellon University. Center for Building Performance and Diagnostics. February 2012.

[25] Construction Industry Council. 2013. *Final Draft Report of the Roadmap for BIM Strategic Implementation in Hong Kong's Construction Industry.* Version 1. September 2013

[26] BuildingSMART. *Industry Foundation Classes Release 4 (FC4).* 2013. Specification Documentation

[27] Liebich, T. 2013. *Summary of IFC Releases. The new buildingSMART Standard*. Presentation for the Official Release Date of buildingSMART's IFC4 – March 12th 2013.

[28] DURAARK. 2014. *Current State of 3D Object Digital Preservation and Gap-Analysis Repor*t. Deliverable 6.6.1. DURAARK Project.

[29] [DURAARK. 2014. Ontological Framework for a Semantic Digital Archive. Deliverable 3.3.2. DURAARK Project.

[30] Karola, A., Lahtela, H., Hänninen, R., Hitchcock, R., Chen, Q., Dajka, S., Hagström, K. 2002. BSPro COM-Server – interoperability between software tools using industrial foundation classes. *Energy and Buildings*. 34 (2002), 901-907.

[31] Pazlar, T. and Turk, Z. 2008. Interoperability in Practice: Geometric Data Exchange Using the IFC Standard. *ITcon*. 13 (2008), 362-380.

[32] Lee, G., Won, J., Ham, S., Shin, Y. 2011. Metrics for Quantifying the Similarities and Differences between IFC Files. *Journal of Computing in Civil Engineering.* 25-2 (2011), 172-1

Supporting the Analysis and Audit of Collaborative OAIS's Using an Outer OAIS-Inner OAIS (OO-IO) Model

Eld Zierau
Department of Digital Preservation
The Royal Library of Denmark
Søren Kierkegaards Plads 1
DK-1016 København K
ph. +45 91324690
elzi@kb.dk

Nancy Y McGovern
Curation and Preservation Services
MIT Libraries
77 Massachusetts Avenue
Cambridge MA 02139 USA
ph. +1 617 253 5664
nancymcg@mit.edu

ABSTRACT

This paper addresses the question: What would distributed digital preservation look like using the OAIS Reference Model? The challenge is the need for several organizations to cooperate to achieve distributed digital preservation; using replication, independence, and coordination to address the known threats to digital content through time. The main purpose of the paper is to present an *Outer OAIS-Inner OAIS* (OO-IO) Model that can support the analysis and audit of collaborative interactions between multiple OAIS's to enable distributed digital preservation. The paper provides extensive explanations and diagrams to demonstrate the ability of the OO-IO model to address distributed digital preservation conformance with the Open Archival Information System (OAIS) Reference Model. It is argued that the OO-IO model contributes a necessary extension to the literature of the digital preservation community to address the analysis and audit necessary for distributed digital preservation.

General Terms

Infrastructure, communities, preservation strategies and workflows, theory of digital preservation, case studies and best practice.

Keywords

OAIS Reference Model, Distributed Digital Preservation, Standards, Audits, Analysis, Collaboration.

1. INTRODUCTION

Digital preservation is the "active management of digital content over time to ensure ongoing access."[1] As good practice for digital preservation matures, organizations are naturally addressing more advanced strategic and operational aspects of the technology required to sustainable digital preservation program leading to distributed digital preservation.

Distributed digital preservation, a focus of this paper, is here defined as "the use of replication, independence, and coordination to address the known threats to digital content through time to ensure their accessibility" ([9] p. 78)[2]. Distributed digital preservation is a form of advanced digital preservation practice, which can be described as in the model for the development of a digital preservation program [4][3]. Here the most advanced stage in that model, externalize, is characterized by collaboration to achieve objectives. In general, it is common in distributed digital preservation for organizations to establish strategic collaborations to meet preservation.

The Open Archival Information System (OAIS) Reference Model is important for digital preservation and is the foundation for the *Outer OAIS-Inner OAIS* (OO-IO) model presented in this paper. The OAIS Reference Model is a core standard in good practice for digital preservation that was approved by the International Standards Organization (ISO) in 2003 and revised in 2012 [1].

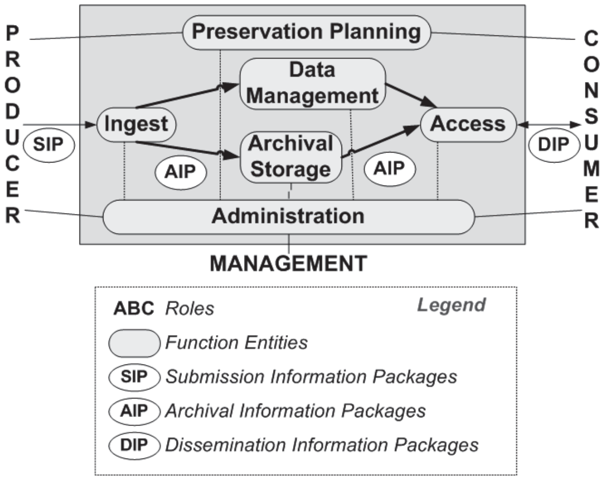


Figure 1. OAIS model

The functional entities and information packages in OAIS Reference Model are depicted in Figure 1, corresponding to

[1] There is no single, authorized definition of digital preservation. The authors cite this definition from the Library of Congress because it is succinct, effective, and often cited in the literature. Available at: http://www.digitalpreservation.gov/about/

[2] Definition is from the *Framework for Applying OAIS to Distributed Digital Preservation* mentioned later in this paper.

[3] The model is discussed in the first chapter of the *Aligning National Approaches to Digital Preservation* (ANADP) volume

Figure 4-1 in the OAIS Reference Model [1]. This will be referred to as a simple OAIS throughout the paper, by contrast to the complexities of the Inner and Outer instances of OASI the paper addresses. OAIS functional entities, functions and information packages will be written in Italic font in this paper.

The OAIS Reference Model provides a framework that has proven effective in guiding the development of sustainable digital preservation programs. "An OAIS is an Archive, consisting of an organization, which may be part of a larger organization, of people and systems that has accepted the responsibility to preserve information and make it available for a Designated Community" (from [1] Section 1-2). References to the term organization in this paper are informed by this definition of an OAIS from the OAIS Reference Model document.

Although the OAIS Reference Model does briefly discuss interoperability for distributed digital preservation in section 6 [1], it needs to be more explicit in order to be usable for analysis and auditing purposes. The OO-IO model can support the analysis and audit of collaborative arrangements between multiple OAIS's, where this paper uses OAIS's as the plural form of an OAIS.

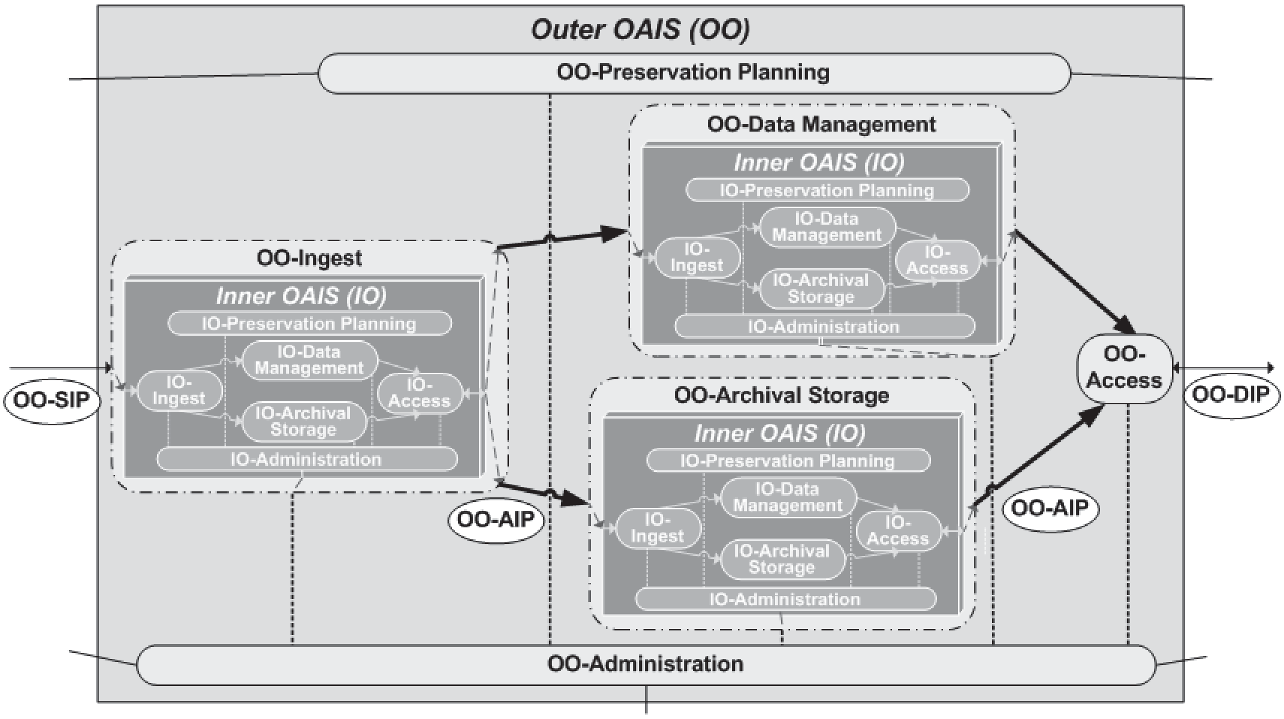


Figure 2. OO-IO model

The OO-IO model is depicted in Figure 2. Explanation of this model will follow later in this paper, as well as how this OO-IO model is building on a previous model (the IR-BR model [8][4]) and the work carried out in the international working group the *Framework for Applying OAIS to Distributed Digital Preservation (DDP)* [9][5],

The *Archival Storage* functional entity of OAIS was the starting point for developing the OO-IO model, just as storage partnerships have been a common starting point for distributed digital preservation. A core requirement of digital preservation is to maintain multiple, geographically-distributed copies of digital content. Meeting that requirement provides a natural opening for storage partnerships and services. The challenge is that the *Archival Storage* needs to be viewed as a distinct OAIS/OAIS's (the *inner OAIS* as the *Archival Storage* is within a separate *outer OAIS*). The reason is that the separated collaboration around *Archival Storage* will need portions of all OAIS functional entities, for example Preservation Planning for media migrations.

Means to support argumentation for conformance to OAIS are needed for distributed digital preservation solutions, which is where the OO-IO model can assist. A decade ago, the majority of organizations in the digital preservation community were focused on determining what it meant to conform to the OAIS Reference Model. The community now includes a growing number of organizations that are engaged in distributed digital preservation. Those organizations have a need to demonstrate conformance with standards through good practice, also for distributed digital preservation.

Section 2 of this paper, as background for the discussion, provides a brief history of the OO-IO model, and places the model into the context of standards and practice for digital preservation, noting developments that informed or led the need for this supplement to further address interoperability in the OAIS reference model. Section 3 explains and illustrates the components of the OO-IO model, and demonstrates the OO-IO model's conformance with a simple implementation of OAIS. Section 4 describes how the OO-

[4] The IR-BR model originates from the pre-study of the Danish BitRepository.org

[5] Preliminary information about the DDP Framework is available at www.metaarchive.org/ddp/index.php/Main_Page.

IO model can support the documentation and audit of collaborative OAIS's.

2. CONTEXT AND NEED

The emergence of good practice for distributed digital preservation that this paper addresses is grounded in the overall context of the development and promulgation of standards and practice for digital preservation. This section traces the development of relevant standards and practice to demonstrate the community-based need for the OO-IO model, as well as the activities that led to its development.

2.1 Standards and Practice

The OO-IO model contributes to the existing foundation of community standards and practice for digital preservation. The model can be used to demonstrate how the complexities of distributed digital preservation use cases can be specified and implemented.

Though digital content has been preserved by some organizations since the 1960s, the digital preservation community traces its roots to the seminal 1996 *Preserving Digital Information report* [7] that defined the problem of digital preservation, specified the challenges that organizations face in managing digital content across generations of technology, considered relevant roles and responsibilities for digital preservation, and framed a set of recommendations to guide the establishment of good practice.

There are several noteworthy things about the 1996 report. The authors of the report represented the domains of the community – libraries, archives, museums, and others – from multiple countries, a rare occurrence at the time, if not a first for the community. The report specified features of digital objects that need to be addressed to ensure the objects' integrity: content, fixity, reference, provenance, and context. In addition, the 1996 report specified the need for the certification of digital repositories that manage digital content to demonstrate good practice and called for "fully distributed storage"[6] of digital objects, a reference to the current challenges of distributed digital preservation. In the nearly twenty years since the *Preserving Digital Information* report was published, a growing set of standards and practice has emerged, as discussed in this section, that provides a frame for good digital preservation practice.

In 1995, the Consultative Committee for Space Data Systems (CCSDS) established the work package that led to the OAIS Reference Model. The OAIS Reference Model references the integrity features, the need for audit and certification of preservation repositories, and the requirement for distributed storage that were specified in the 1996 report [7]. OAIS was approved as an international standard in 2003 and updated in 2012. Most organizations that are responsible for the long-term preservation of digital content refer to and in many cases build and implement their repository systems to align with the principles and concepts of OAIS. These activities demonstrate that OAIS is being maintained and is in use within a significant portion of the community, two measures of success for standards that have a demonstrated impact on practice.

The OAIS Reference Model is not a standalone standard, but the anchor for a family of OAIS-related standards. One of the characteristics that has enabled OAIS to endure is the standards roadmap that has been included since the early drafts of the document that addresses the ways in which OAIS needs to be extended and applied.[7] Examples of standards that are called for in the OAIS standards road map and that are cited the 2012 update of the Reference Model include: the Producer-Archive Interface Method Abstract Standard (PAIMAS [3]); and preservation metadata, e.g., Preservation Metadata Implementation Strategies (PREMIS [5]).

Another standard in the OAIS family addresses the need for audit and certification to enable digital repositories to demonstrate good practice. Audit and Certification of Trustworthy Digital Repositories (ISO-16363 [2]) is a standard that was built on the Trustworthy Repository Audit and Certification (TRAC) requirements [6]. The audit and certification requirements for digital preservation stipulate that organizations provide evidence to demonstrate how they conform to the ISO 16363 requirements.

Demonstrating good practice for distributed digital preservation is complicated by the need to accumulate and consolidate evidence across collaborating OAIS's. The OO-IO model supports audit requirements within distributed digital preservation environments by elaborating the relationships and roles of functional entities and their functions within and between relevant OAIS's.

Section 6 of the OAIS Reference Model serves as a reference point for the OO-IO model within the current framework of digital preservation standards, That discussion in OAIS considers issues pertaining to interoperability between archives and levels of interaction between OAIS Archives (Section 6-1) and Management issues with federated archives (Section 6-2). This portion of OAIS acknowledges the need for interoperability in digital preservation, but the discussion is not extensive and does not specify an approach for achieving interoperability. Practitioners of distributed digital preservation are developing an understanding of how interoperability can be realized [9]. The methodology of the OO-IO model, informed by that deepening understanding, involved systematic analyses of common use cases for distributed digital preservation that are described in Section 2.2 and elaborated in Section 3.

2.2 Provenance of the OO-IO model

The development of the OO-IO model was initially motivated by the complexities of good practice for distributed digital preservation that were identified by organizations that have become engaged in distributed digital preservation. In practice, distributed digital preservation involves a range of use cases to address the specifics of interoperability between multiple OAIS's.

It was an investigation of such complexities that led to the development of the Institution Repository–Bit repository (IR-BR) model [8], the starting point for the OO-IO model. The IR-BR model emerged during work on the open source BitRepository.org framework that is used for bit preservation in Danish Cultural institution. Later in this paper, the correlations between the IR-BR model and the *Archival Storage* component of the OO-IO model are explained. In the IR-BR model, the Institution Repository is an Outer OAIS as an organization using a Bit Repository that is an Inner OAIS.

[6] Citation from the 1996 Report [7].

[7] OAIS section 1.5.

The IR-BR model informed and influenced the development of the *Framework for Applying OAIS to Distributed Digital Preservation (DDP)* [9][8], a result of a project established to address the growing awareness of the need to adapt and extend current standards to address distributed digital preservation, models and auditing methodologies to support DDP. The DDP Framework addresses the roles, functions, and use cases that build a layer upon section 6 of the OAIS Reference Model to begin to specify how interoperability and federation might work. The DDP Framework has been developed by a working group with representatives from both North America and Europe that included the authors of this paper and representatives from some major DDP examples, including MetaArchive, the Danish BitRepositorty.org, Chronopolis, Data-PASS, DuraCloud, Internet Archive, UC3 Merritt and Archivematica. Variations within this range of cases pointed to the need to focus on other OAIS functional entities that require distribution over more organizations requiring a generalization of the IR-BR model into the OO-IO model. The results of the DDP Framework project will be shared when available.

Developing the OO-IO model provided the means to analyze the functionality of an inner OAIS and provided common terminology between inner and outer OAIS's. The generalization in the OO-IO model applies not only to the *Archival Storage* functional entity that can be seen as a separate Inner OAIS, but also *Data Management* and *Ingest*. The following section explains and demonstrates the feasibility and validity of this generalization.

3. THE OUTER OAIS-INNER OAIS MODEL

The primary purpose of the Outer OAIS–Inner OAIS (OO-IO) model is to simplify the challenges – organizational (what needs to be done) and technological (how it can be done) - of engaging in distributed digital preservation that involves several organizations. An Outer OAIS refers to an entire OAIS implementation – a simple OAIS – that supports distributed digital preservation, including all of its Inner OAIS's. An Inner OAIS is an OAIS that is distinct from the Outer OAIS and is implemented to manage one OAIS functional entity - *Ingest*, *Data Management* or *Archival Storage*. Each inner OAIS is managed as a complete OAIS, though it is dedicated to managing a single functional entity in the Outer OAIS, as depicted in Figure 2. One example of a case that requires the OO-IO model rather than a simple OAIS is when the functional entity (e.g., *Archival Storage* as a bit repository) is separated and managed by one or more external organizations (OAIS's), as is often the case in distributed digital preservation.

Note that the sample *Inner OAIS* cases that have been specified in this paper for the OO-IO model (i.e. *Archival Storage*, *Ingest*, and *Data Management*) focus on functional entities that require storage because an inner OAIS without storage would not be necessary. The functional entities that require storage are those that interact directly with *SIPs*, *AIPs* and *DIPs*. These information packages are pictured in Figure 1. That figure illustrates Submission Information Packages (*SIPs*) being received via the functional entity *Ingest*. The *Ingest* functional entity then creates Archival Information Packages (*AIPs*) and the related data management information that are parsed to the *Archival Storage* and *Data Management* functional entities, respectively. In response to an *Access* request for Dissemination Information Packages (*DIPs*), the *AIPs* and related data management information required to create the *DIP* are delivered via the *Access* functional entity. There is no sample case for the *Access* functional entity in the OO-IO model because *Access* generates/re-generates *DIPs* based on information received the *Data Management* and the *Archival Storage* functional entities. Thus there are no obvious cases for risk of loss or need of cooperation in relation to the *Access* functional entity

The 'archive interoperability' discussion in Chapter 6 of the OAIS Reference Model states that an OAIS may be geographically distributed. It lists possibilities of all components being under the same *Management*, or spread over OAIS Archives with separate *Managements* that work cooperatively. The OO-IO model builds upon Chapter 6 and in doing so, the elaboration of the OO-IO model aligns with the existing OAIS reference model. The OO-IO model specifies an approach for using the OAIS model to achieve archive interoperability that Chapter 6 does not provide.

A strength of the OO-IO model is that the analysis required to develop the model demonstrates the need for the parsing of OO functional entities into OO and IO functional entities, as does the analysis of the interface between the OO and the IO. The systematic process for developing the OO-IO model identified the prefix for terms (OO or IO), making clear distinctions between inner or outer OAIS functions and information. This specification verifies that the OO-IO model conforms to the OAIS Reference Model. Therefore, the inner OAIS scenario is detailed to demonstrate the case for *Archival Storage*, *Ingest* and *Data Management* in the below sections.

3.1 The OO-IO *Archival Storage* Component

For distributed digital preservation, one use case for the *Archival Storage* component of the OO-IO model is the need to operate a separate standalone bit repository to meet the requirements of the *Archival Storage* functional entity of the *Outer OAIS*. The standalone bit repository itself is managed as an *Inner OAIS* and incorporates some of all of the functional entities of an OAIS.

The *Archival Storage* component of the OO-IO model addresses only the *Archival Storage* functional entity of the OAIS Reference Model. It is an inner OAIS as depicted in Figure 3.

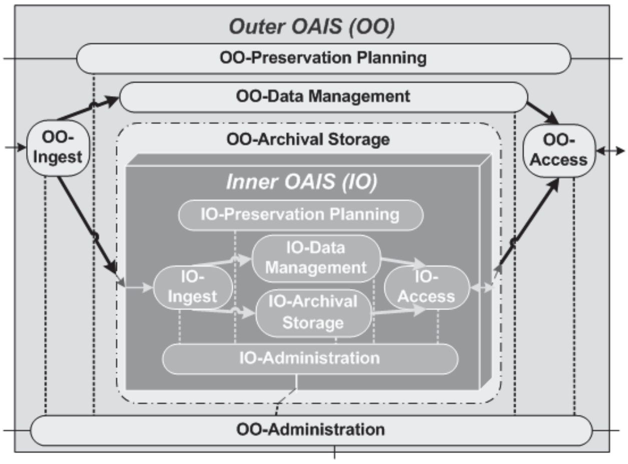


Figure 3. Archival Storage in the OO-IO model

[8] Preliminary information about the DDP Framework is available at www.metaarchive.org/ddp/index.php/Main_Page.

Distributed digital preservation implementations of OAIS require more OAIS functions, e.g., *Ingest* including the *Receive Submission* function, in addition to *Archival Storage* functions (depicted and described later in Figure 4[9]). The bit repository must be treated as an Inner OAIS where parts of *all* of the OAIS functional entities are required by the inner OAIS.

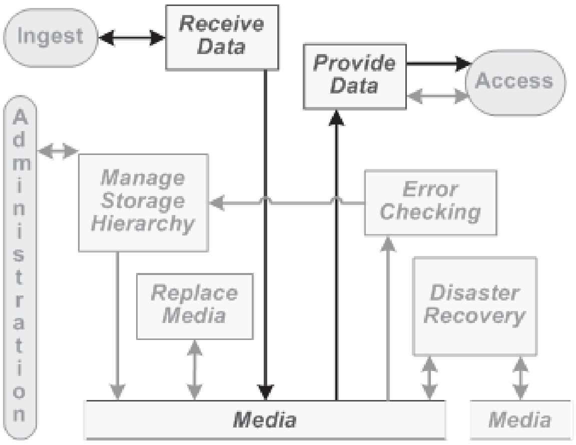


Figure 4. Functions of OO-IO *Archival Storage*

3.1.1 Flow for the Archival Storage component

In developing the OO-IO model, an investigation of the flow of information from *Ingest* to *Access* of the Outer OAIS confirmed the validity and utility of the *Archival Storage* component of the OO-IO model.

Figure 5 illustrates a high-level flow of the information packages with focus on the *Archival Storage* component of the OO-IO model. The dotted lines in the figure indicate that there are more functions involved in the path.

In this flow, an Outer OAIS *Submission Information Package* (*OO-SIP)*[10] is received from an *OO-PRODUCER*[11] and passed to the *OO-Ingest* functional entity. All of the internal *OO-Ingest* functions are executed, resulting in the transformation of *OO-SIPs* to *OO-AIPs*. The difference for Archival Storage in the OO-IO model occurs during the transfer of an Outer OAIS *Archival Information Package* (*OO-AIP)* to the *OO-Archival Storage.*

When the *OO-AIP* is transferred to *OO-Archival Storage,* it takes an alternative path from a simple OAIS implementation when it is ingested into the Inner OAIS within the *OO-Archival Storage*. Thus an *OO-AIP* becomes an *IO-SIP* and runs via the *IO-Ingest* functions before it is transformed into an *IO-AIP* in the *IO-Archival Storage*. Likewise, the receipt/storage confirmation for accepted data and completed storage is returned to the Outer OAIS from the *IO-Ingest*, as the inner OAIS acts as an OAIS.[12]

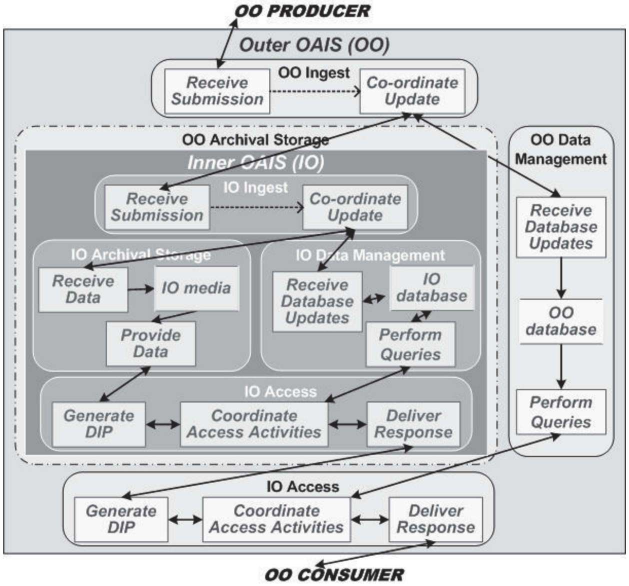


Figure 5. Information path for *Archival Storage* component

3.2 The OO-IO *Ingest* Component

The *Ingest* component of the OO-IO model addresses only the *Ingest* functional entity of OAIS, It is an Inner OAIS as depicted in Figure 6.

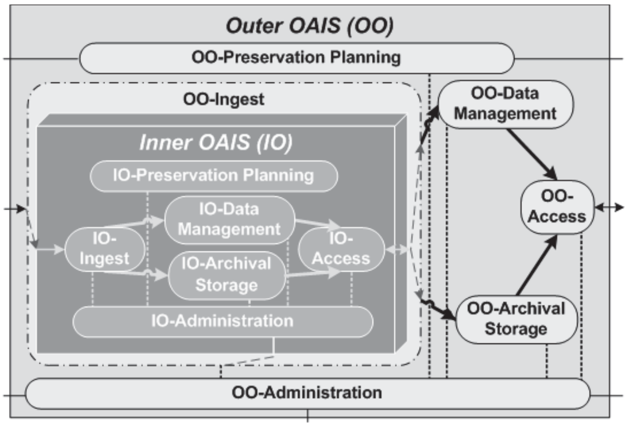


Figure 6. *Ingest* component of the OO-IO model

There are two use cases that demonstrate the need for Ingest as an Inner OAIS: distributed ingest and delayed processing in ingest, as discussed below.

Distributed ingest is a scenario that was identified in several cases developed for the DDP project. In particular, micro-service-based solutions like UC3-Merritt and Archivematica had examples of

[9] From the OAIS Reference Model Figure 4-3: Functions of the Archival Storage Functional Entity [1].

[10] Submission Information Package – see Figure 1.

[11] OAIS roles appear in all capitals in this paper.

[12] A similar argument can be made for the Access component. The full explanation and supporting justification can be found in the "Cross Institutional Cooperation on a Shared Bit Repository" [8]. This also includes examples of a possible split of OO-Data Management and IO-Data Management, OO-Administration and IO-Administration and OO-Preservation Planning and IO-Preservation Planning.

using the distribution of micro-services to manage many simultaneous loads of ingest processing.[13]

Preliminary archiving of *SIPs*, is a scenario where *SIPs* are secured in order to mitigate risk of losing them due to risk of loss caused by delays in the ingest process. The Royal Library of Denmark has focused on this as a result of a general risk analysis completed for the digital preservation program at the library. One of the reasons for a delay in archiving can be that it takes time to get all the information needed to generate *AIPs*. For instance a computer game may need trailers and digitized user guides before it can be archived. Another reason can be that large digitization project may have interdependent data that needs to be connected before archiving can proceed. A third reason can be that digital material can require a lot of work before becoming an *AIP*, for instance hard drives from deceased authors, which must be analyzed and restructured before becoming an *AIP*.

Like *Archival Storage*, the *Ingest* component of the OO-IO model is an Inner OAIS that functions as a separate ingest mechanism, including its own *Archival Storage*, within the *OO-Ingest*. The Inner OAIS must also include portions of all of the OAIS functional entities, not only Ingest.

3.2.1 Information flow for the Ingest component

In developing the OO-IO model, an investigation of the flow of information from *Ingest* to *Access* of the Outer OAIS demonstrated the validity and utility of the *Ingest* use case of the OO-IO model.

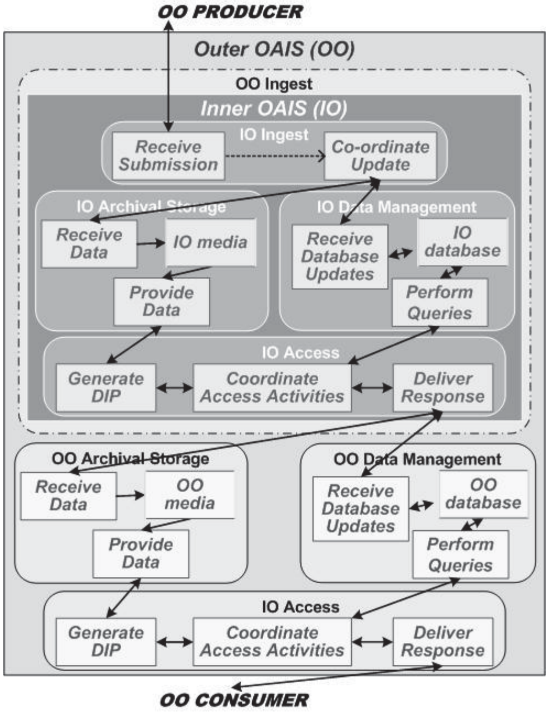


Figure 7. Information path for *Ingest* component

Figure 7 illustrates the information flow for *Ingest*, in the same way that Figure 5 did for *Archival Storage*. The flow is somewhat more complex for *Ingest* because this functional entity delivers information to both *Archival Storage* and *Data Management*.

In this flow, an *OO-SIP* is received from an *OO-PRODUCER* and passed to the *OO-Ingest* functional entity. Already there is a change because the *OO-SIP* takes alternate path by being ingested to the *IO* instead of being passed to *OO-Ingest* functions. An *OO-SIP* becomes an *IO-SIP* and runs via the *IO-Ingest* functions becoming an *IO-AIP*, which are secured in *IO-Archival Storage*. A closer look at the *Ingest* functions in Figure 8 makes this clearer.

The *Ingest* component of the OO-IO is also more complex than the *Archival Storage* component of the OO-IO because the *Ingest* component is ***not*** only a question of the information taking another path before reaching a destination (like the *OO-AIP* taking another path before reaching the *IO-Archival Storage)*. In the *Ingest* component, it is only the *IO-SIP*/*OO-SIP* that is sent to *IO-Archival Storage*, which means that *Ingest* functions corresponding to *OO-generate AIP* and *OO-Coordinate Update* are not expressed in the IO as they are in the OO. This means that the *OO-Ingest* must generate an *OO-AIP* and coordinate updates, while only the *OO-SIP* (possibly with a minimum of metadata) is secured in the *IO-Archival Storage*. The functions performed within the *IO-Access* functional entity are to generate the *OO-AIP* and coordinate updates.

Viewed from this perspective, it makes perfect sense that the *IO-DIP* can be associated with an *OO-AIP*, since a *DIP* is derived from an AIP to fit the request from a *CONSUMER*. For the *Ingest* component of the OO-IO model, an *IO-CONSUMER* (the *OO-Archival Storage/OO-Data Management*) gets an *IO-DIP* (or rather an *OO-AIP)* that is derived information from an *IO-AIP*. The *IO-Ingest* takes the *OO-SIPs* as *IO-SIPs* and transfers them to the *IO-Archival Storage* without transformations (although some minimum information may be added). This makes the *IO-AIP* equivalent (or very similar) to the *IO-SIP/OO-SIP*. This requires the *IO-Access* functional entity to operate like the *OO-generate AIP* and *OO-Coordinate Update* of an ordinary *OO-Ingest* functional entity.

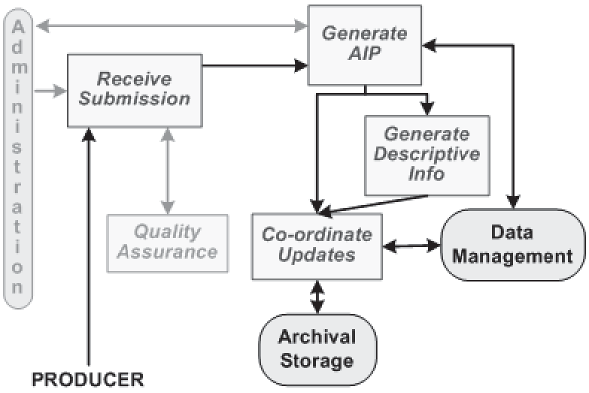


Figure 8. Functions of the *Ingest* functional entity

To further explain this portion of the analysis, the functions of the *IO-Access* functional entity are depicted in Figure 9.

[13] See UC3-Merritt at: https://merritt.cdlib.org/ and Archivematica at: https://www.archivematica.org/wiki/Main_Page.

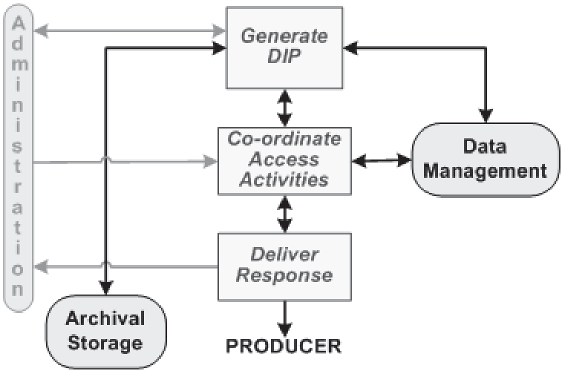


Figure 9. Functions of the *Access* functional entity

The *Ingest* functions that should correspond to *OO-Generate AIP* (and *OO-Generate Descriptive Info*) need to be included in the *Generate DIP function* of the *IO-Access* functional entity. As the redrawing[14] of the functions in Figure 8 and 9 show, there is similar flow through functions with similar names and meaning:

- the *PRODUCER* in Figure 8 matches the *Archival Storage* of Figure 9,
- the *Generate AIP* (together with *Generate Descriptive Info*) in Figure 8 matches *Generate DIP* of Figure 9,
- the *Archival Storage* in Figure 8 matches the *CONSUMER* of Figure 9, and
- the *Coordinate Updates* Figure 8 aligns with *coordinate Access Activities* of Figure 9.

In practice, it will be important to pay close attention to how this portion of the expected *OO-Ingest* functions map to these *IO-Access* functions.

For *Preservation Planning* in the *Ingest* component of the *OO-IO* model, the *IO-Preservation Planning* is only concerned with security of the ingested *IO-SIP* (corresponding to the *OO-SIP*). This means that *Preservation Planning* is split between the *IO* and the *OO-Preservation Planning* where *OO-Preservation Planning* covers all other *Preservation Planning* for the *OO-SIPs*. This may require coordination as in the example of the *Archival Storage* component of the OO-IO model.

For *MANAGEMENT* in the *Ingest* component of the OO-IO model, it is generally true – as it was for the *Archival Storage* component of the OO model – that requirements resulting in directions from *OO-MANAGEMENT* are dealt with by *IO-Administration*. From the *IO* perspective, the *OO-Administration* represents *IO-MANAGEMENT*. It is at the interface between *OO-Administration* and *IO-Administration* that the mapping of the requirements for the *IO* takes place.

For *Data Management* in the *Ingest* component of the OO-IO model, there may be specific *IO-Data Management* actions that are only relevant to the *IO*, but there will most likely also be elements of *IO-Data Management* that must be passed to the *OO-Data Management*. Examples include catalogs, inventories and audit trails. This portion of the OO-IO model works well as the *IO-DIP* (that becomes the *OO-AIP*, updating *OO-Archival Storage* and *OO-Data Management*) is generated from *IO-Archival Storage* as well as from the *IO-Data Management.*

3.3 The OO-IO *Data Management* component

The *Data Management* component of the OO-IO model addresses only the *Data Management* functional entity of the OAIS Reference Model. It is an Inner OAIS as depicted in Figure 10.

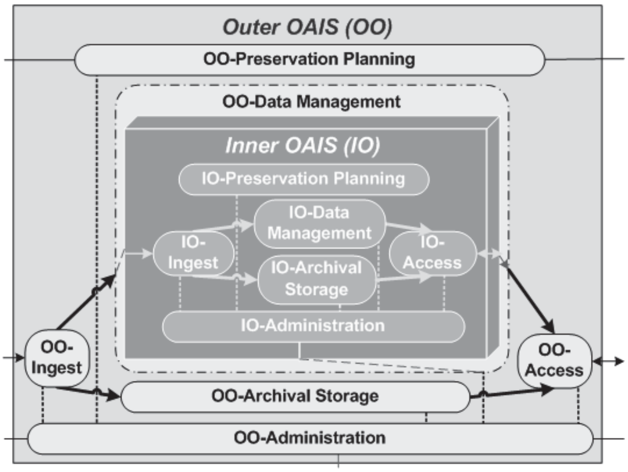


Figure 10. *Data Management* component of OO-IO model

There may be different use case scenarios, where it can make sense to have *Data Management* as an inner OAIS. The following scenarios are just examples:

- *Separate securing of data.* A situation that addresses multiple instances of content that has security requirements.
- *Distributed linked data representing database* A case where a database is represented by linked data that is distributed across multiple environments.

Separate secure data is a scenario similar to the one for *Ingest*. There may be portions of *Data Management* data that need to be secured for distributed digital preservation. This can occur when there is a need for an asynchronous update of *Data Management* in connection with ingests of digital materials or for the ongoing creation of collection information that later may be needed for preservation.

Distributed linked data represented in the database is a scenario where linked data are represented in the *Data Management*, implying that *Data Management* is distributed. This case could require descriptive elements from databases that are managed by different organizations. This is especially relevant for representation information, if for instance:

- One organization has descriptions of its preserved assets
- Another organization has the format registry used for the preserved assets
- A third organization has the environment registry used for the preserved assets

Distributed knowledge-bases like registries are individually maintained, and it make sense to have separate mechanisms for

[14] Simplification of duplicated arrows and moving the entities, functions and roles around compared to the illustrations in the OAIS Reference Model

preservation planning, e.g., policies for maintenance of registries that are used as shared resources.

Like *Ingest*, the *Data Management* component of the OO-IO model is an Inner OAIS that functions as a separate data management mechanism, including its own *Archival Storage* as well as portions of all the functional entities of an OAIS.

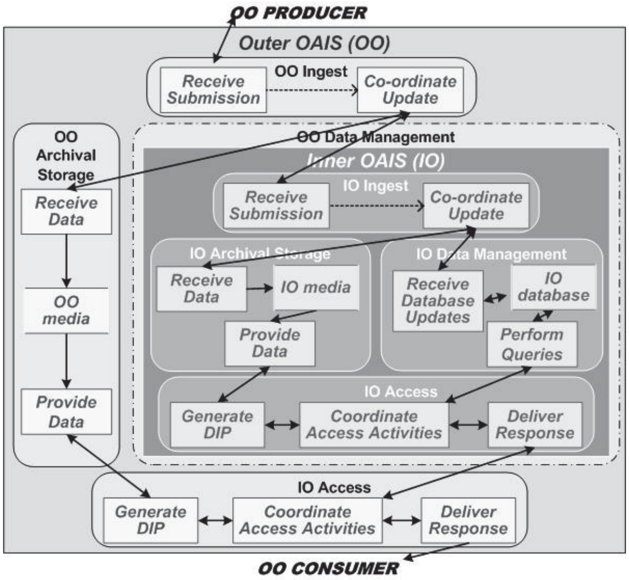


Figure 11. Flow for *Data Management*

3.3.1Flow for the Data Management component

In developing the OO-IO model, an investigation of the flow of information from *Ingest* to *Access* of the Outer OAIS validated the *Data Management* component of the OO-IO model. Figure 11 illustrates the information flow, as Figure 5 did for *Archival Storage*.

In the *Data Management* flow, an *OO-SIP* is received from an *OO-PRODUCER* and passed to the *OO-Ingest* functional entity. All the internal *OO-Ingest* functions are executed, transforming *OO-SIPs* to *OO-AIPs* and belonging to *OO-Data Management* information. The change occurs when *OO-Data Management* information, e.g., reports and update information, is transferred to the *OO-Data Management* because it takes an alternate path, being ingested into the Inner OAIS instead of being passed to *OO-Data Management* functions. This data management information becomes an *IO-SIP* and runs via the *IO-Ingest* functions before it ends as an *IO-AIP* in the *IO-Archival Storage.* Here, the *IO-Archival Storage* containing the *IO-AIP* (*OO-Data Management* information) may be seen as equivalent to the *OO-Data Management* database. The *IO-Access* acts as the *Perform Queries* function of the *OO-Data Management* functional entity.

Like the *Ingest* component of the OO-IO, the *Data Management* component of the OO-IO model is more complex than the *Archival Storage* component. Similarly, this is because the *OO-Data Management* functions – in a simple OAIS implementation - are not just taken over by the *IO-Data Management* functions, but have to be interpreted in terms of other IO-functional entities and functions. However, the *Data Management* component is simpler than the *Ingest* component because the *Ingest* and *Access* information for the functions of *Data Management* are more similar to OAIS, than the *Ingest* and *Access* information for *Ingest*.

Administration functions are managed within the OO and the IO. However, *OO-Administration* report requests from *OO-Data Management* can be regarded as either a report request to the IO from *IO-MANAGEMENT* or from *IO-Access* (if *IO-DIPs* are considered to be a report result). It may also be a mix of these depending on the type of reports requested.

As with the *Ingest* use case, *Preservation Planning* for the *Data Management* use case is split between the OO and the IO. *OO-Preservation Planning*, among other things, covers the function *Develop Preservation Strategies and Standards.* For the *Data Management* component of the OO-IO model, the split follows the split of responsibilities. For example, a format registry in the IO will include the *Develop Preservation Strategies and Standards* function for this registry, while other functions of *Preservation Planning*, like the *Monitor Technology* function that addresses issues like media for *Archival Storage*, are included in the OO.

Also for the *Data Management* component of the OO-IO model, all requirements resulting in directions from *OO-MANAGEMENT* are dealt with by *IO-Administration*, as described in the end of the previous section on the OO-IO *Ingest* component.

4. USING THE OO-IO MODEL

There is a range of use cases for which the OO-IO model can be advantageous for distributed digital preservation.

First, the OO-IO model provides a means to explicitly express the OAIS functional entities and functions that are referred to by prefixing then with OO (for Outer OAIS) and IO (for Inner OAIS) for each component of the OO-IO model. Although this may seem trivial, the experience from the Danish use of the model is that it can improve communications. The use of Inner and Outer qualifiers for discussions that involve distributed digital preservation can avoid misunderstandings.

Second, the OO-IO model provides a basis for analysis of the interfaces between an *Outer OAIS* and an *Inner OAIS* that are essentially for understanding and implementing interoperability that is essential for distributed digital preservation. The OO-IO model diagrams make it be possible to explicitly map inputs and outputs that inform or produce required evidence for audit. In using the Archival Storage component of the OO-IO model, this analysis has proven to be extremely useful, both initially for the design and later the auditing of the Danish BitRepository.org.

Third, the OO-IO model can support and enable audit for distributed digital preservation. The development of the OO-IO model produced a generalized model that addresses distributed digital preservation and is grounded in standards and practice. Though it can and should be extended, this version of the model can provide a framework self-assessment and audit processes for distributed digital preservation.

A challenge for audit within distributed digital preservation environments is mapping responsibilities and accumulating evidence across multiple OAIS's to cumulatively demonstrate compliance with digital preservation requirements as specified in ISO 16363. The OO-IO model supports audit for distributed digital preservation:

- By allowing the paths (roles, functions, inputs, and outputs) between Outer and Inner OAIS's to be mapped,
- By providing a framework, based on that mapping, to determine which components of Inner OAIS's and Outer OAIS's address specific ISO 16363 requirements, and
- By directly supporting the completion of a gap analysis, using the ISO 16363 requirements, in preparation for an audit (peer review or external) of a distributed digital preservation environment.

Summing up the OO-IO model can support the analysis and audit of collaborative interactions between multiple OAIS's to enable distributed digital preservation. This section has highlighted the benefits of the OO-IO model to improve communication, for developing and managing Inner and Outer OAIS's, and for supporting the audit of collaborative OAIS's

5. DISCUSSION AND FURTHER WORK

A challenge, though not insurmountable, in using and applying the OO-IO model is the complexity of the cases that detail the roles, functions, interactions, and outcomes of the interoperability between and within OAIS's that is required to manage distributed digital preservation environments. Therefore, working with the OO-IO model requires a deeper familiarity with and understanding of the workings of OAIS than is required for more simple use cases and implementations.

The different components of the OO-IO model have varying degree of complexity. For example, the *Ingest* component introduces an additional complexity by defining results from the *OO-Ingest* as the result of *IO-Access* of the Inner OAIS, i.e. the product *OO-AIP* for the Outer OAIS is part of the *IO-DIP* of the *Inner OAIS*, but also the *OO-Data Management i*nformation is part of this *IO-DIP*.

The example of using the OO-IO model to support and enable audit for distributed digital preservation also highlights further work that is needed on the model to elaborate use cases that illustrate and document audit processes. The productive discussions that occurred in DDP cases while developing the DDP Framework suggest that:

- an increasing number of practitioners are interested in and need to use DDP use cases,
- DDP cases contribute timely implementation examples to the literature of the digital preservation community, and
- DDP cases provide examples that can be used for academic and continuing educational purposes.

Now that the OO-IO model and the DDP Framework have been specified and both will be shared with the community, the OO-IO model and the DDP framework would benefit from an evaluation by more DDP organizations and by the broader digital preservation community.

Although the paper makes the case that the OO-IO model only makes sense for OAIS functional entities that involve storage of information packages, use cases may emerge that indicate the need to extend the model to also address *Preservation Planning* and *Administration*. These entities do not have obvious cases, because these functional entities do not come into direct contact with information packages in an OAIS. However, these functional entities could be investigated further and added to the OO-IO model, if relevant cases arise. The same applies for the *Access* functional entity if relevant cases arise.

6. CONCLUSION

In summary, the *Outer OAIS-Inner OAIS* (OO-IO) Model is needed to support the specification and audit of collaborative interactions between multiple OAIS implementations for distributed digital preservation.

The paper has provided extensive explanations and diagrams to make evident the ability of the OO-IO model to address distributed digital preservation conformance with the OAIS Reference Model.

The need for and utility of the OO-IO model as a supplement to the literature documenting current standards and practice for digital preservation was discussed then demonstrated using a sample of use cases for distributed digital preservation.

7. ACKNOWLEDGMENTS

We would like to thank participants in the DDP Framework group, since the evolved OO-IO model is based on work done in this group. We would also like to thank Sébastien Peyrard from The National Library of France, Kari René Smith from The Massachusetts Institute of Technology, and Ulla Bøgvad Kejser from The Royal Library of Denmark for useful input and discussions.

8. REFERENCES

[1] ISO 14721:2012. 2012. *Space data and information transfer systems - Open archival information system (OAIS) - Reference*. Retrievable via http://www.iso.org/iso/catalogue_ics.

[2] ISO 16363:2012. 2012. *Space data and information transfer systems - Audit and certification of trustworthy digital repositories*. Retrievable via http://www.iso.org/iso/catalogue_ics.

[3] ISO 20652:2006. 2006. *Space data and information transfer systems -- Producer-archive interface -- Methodology abstract standard*. Retrievable via http://www.iso.org/iso/catalogue_ics.

[4] McGovern, N.Y., Skinner, K. (Editors). 2012. *Aligning National Approaches to Digital Preservation*. Publisher: Educopia Institute Publications, Atlanta, Georgia. Available at http://metaarchive.org/public/publishing/Aligning_National_Approaches_to_Digital_Preservation.pdf, retrieved March 2014.

[5] *Preservation Metadata: Implementation Strategies (PREMIS)*, hosted at http://www.loc.gov/standards/premis/. Retrieved March 2014

[6] RLG/NARA Task Force on Digital Repository Certification. 2007. *Trustworthy Repositories Audit & Certification: Criteria and Checklist*. Publisher: Chicago, CRL.

[7] Waters, D., Garrett, J. 1996. *Preserving Digital Information: Report of the Task Force on Archiving of Digital Information*. Available at http://www.clir.org/pubs/reports/pub63watersgarrett.pdf, retrieved March 2014

[8] Zierau, E., Kejser, U.B. 2013. *Cross Institutional Cooperation on a Shared Bit Repository*, In: Journal of the World Digital Libraries, vol. 6, no. 1.

[9] Zierau, E., Schultz, M. 2013. *Creating a Framework for Applying OAIS to Distributed Digital Preservation*, In: Proceedings of the 10th International Conference on Preservation of Digital Objects, Lisbon, Portugal, 2013.

Identifying Digital Preservation Requirements: Digital Preservation Strategy and Collection Profiling at the British Library

Michael Day
The British Library
96 Euston Road, London NW1 2DB
United Kingdom
+44 (0)843 2081144 x 3364
Michael.Day@bl.uk

Ann MacDonald
University of Kent
Canterbury, Kent, CT2 7NZ
United Kingdom

Akiko Kimura
The British Library
96 Euston Road, London NW1 2DB
United Kingdom
+44 (0)20 7412 7214
Akiko.Kimura@bl.uk

Maureen Pennock
The British Library
Boston Spa, Wetherby
West Yorkshire LS23 7BQ
United Kingdom
+44 (0)1937 546302
Maureen.Pennock@bl.uk

ABSTRACT

The British Library is increasingly a digital library. Over past decades, it has built up significant collections of digital content covering a very wide range of content types. In addition to the increasing amounts of digital content acquired by purchase or donation, the Library and its partners have also invested heavily in the digitization of selected collection content, helping to create large collections of certain types of content (e.g., newspapers, out-of-copyright books, and sound). Most recently, the extension of legal deposit provisions to non-print works in 2013 has meant that the British Library - working in conjunction with the other UK legal deposit libraries - has begun to collect new categories of digital content, including periodic harvests of the UK Web domain. In order to support this, the Library has also invested heavily in developing scalable infrastructures for the acquisition, storage and management of large amounts of digital content. The British Library Digital Preservation Strategy, 2013-2016 is focused on the embedding of digital sustainability as an organizational principle across the Library and to help manage preservation risks and challenges across all digital collection content lifecycles. This practice paper describes work being undertaken by the Digital Preservation Team at the British Library to develop content profiles of high-level digital collections that will support the implementation of the strategy, in particular for the capture of long-term preservation requirements.

General Terms

strategic environment, preservation strategies and workflows, case studies and best practice

Keywords

digital preservation, collection content profiling, preservation planning, institutional contexts of preservation

1. INTRODUCTION

This paper describes work being undertaken by the Digital Preservation Team at the British Library to develop a content profiling framework for high-level digital collections that will help support the capture of long-term preservation requirements. The resulting collection profiles are short human-readable documents that document and contextualize collections that then can be used as part of the preservation planning process.

This paper will follow the following structure. After a section describing the digital preservation context of the British Library, section 3 will outline related work in the areas of preservation planning, content characterization and profiling, the capture of preservation intent, and some approaches to institution-level assessment. Section 4 will then describe in more detail: 1) challenges around the identification of high-level digital collections at the British Library, and 2) the development of the initial collection profile framework. Section 5 provides some conclusions and pointers to future work.

2. THE BRITISH LIBRARY CONTEXT

The British Library is the UK's national library; its role is defined in legislation as "a national centre for reference, study and bibliographical and other information services, in relation both to scientific and technological matters and to the humanities" [British Library Act 1972].

2.1 Legal Deposit

As a legal deposit library, the British Library has the right to receive a copy of printed content published in the UK (including books, newspapers, printed music and maps) as well as - since April 2013 - certain kinds of non-print content. For printed materials, this obligation has existed in English law since the seventeenth century. Primary legislation supporting the extension of legal deposit to non-print items in the UK was passed in 2003.

After a decade of planning and negotiation, official regulations came into force on the 6th April 2013 [Legal Deposit Libraries (Non-Print Works) Regulations 2013]. This, for the first time, enabled the British Library and the other UK copyright libraries to claim certain classes of non-print content under legal deposit provisions and make it available to on-site users [Gibby and Brazier 2012].

This has included the scaling-up of the Library’s existing Web archiving activities to include a periodic capture of the entire UK Web Domain, the first of which (running from April to June 2013) captured 31TB of compressed data [Webster 2013]. It has also led to the development of specialised ingest workflows for the capture of other kinds of published content, including e-journals and e-books.

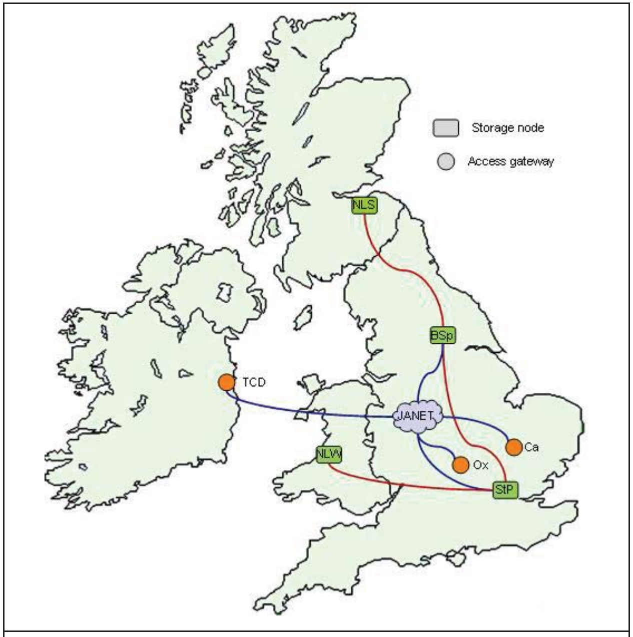
2.2 Infrastructures

In order to scale-up its technical infrastructure, the British Library and the other UK Legal Deposit Libraries have invested heavily in developing scalable solutions to the acquisition, storage and management of very large amounts of digital content. The resulting Digital Library System (DLS) has been described as a “single location to ingest, store, preserve, manage, discover and provide controlled access to digital content assets” [Fleming 2011]. While designed as an integrated storage system, it has been implemented in a highly distributed way, with content replicated in four storage nodes (based in London, Boston Spa, Edinburgh and Aberystwyth) with additional access gateways at the university-based legal deposit libraries (Figure 1).

Some features of DLS have been described in an APARSEN project deliverable [APARSEN 2013]. Ingest takes place at either of the British Library’s sites, with different ingest streams defined for different types of content, e.g. e-journals, digitized newspapers, or web archive content. All objects have a signature file, which includes a hash value and timestamp, and content is automatically replicated on all four storage nodes after ingest. The system assumes that in a large-scale storage system, some bit-loss is inevitable. DLS has, therefore, been designed to be self-checking and self-healing; there are periodic integrity checks, and “if an object is found to be damaged, it is replaced by a good copy from another node” [APARSEN 2013]. DLS is designed to be scalable and vendor-independent, using commodity hardware that can be added to as required.

2.3 Strategy

At the same time, the Library has begun to try to understand what might be meant by a “national collection” in a digital age. It has been widely recognized that the meaning of traditional concepts of “collection” (and therefore “collections management”) have changed significantly in the digital era, e.g. being focused much less on ‘tangible’ content held and managed locally and more on providing access to content held elsewhere [Corrall 2011; Corrall and Roberts 2012]. In this environment, a great deal of attention needs to be given to access rights. For example, Brazier has commented that “access rights are replacing physical ownership as the fundamental definition of being ‘in’ a library collection” [Brazier 2013]. This shift is also seen in the British Library’s Content Strategy, 2013-2015. While recognizing the continuing significance of collecting activity, e.g. through legal deposit, voluntary deposit and donation, the strategy states that outside of this, “the Library will prefer to connect to content, except in circumstances where the connection is not technically feasible or when we wish to hold and preserve the materials for the long term” [British Library 2013a]. Despite this, the logic of Non Print Legal Deposit, Web domain harvesting, and the Library’s ongoing digitisation partnerships mean that the amount of digital content that will require long-term preservation is growing at an extremely rapid rate.



Storage nodes:

- British Library, St Pancras (STP)
- British Library, Boston Spa (BSP)
- National Library of Wales (NLW)
- National Library of Scotland (NLS)

Access gateways:

- Bodleian Library, Oxford (Ox)
- Cambridge University Library (Ca)
- Trinity College Library, Dublin (TCD)

Figure 1. DLS Storage Nodes (Source: APARSEN 2013)

When all of this is taken into account, it is clear that the British Library is increasingly becoming a digital library. The British Library’s Digital Preservation Strategy, 2013-2016 starts from the assumption that it is the Library’s responsibility to preserve and make available this content to current and future users, while noting, however, that “preservation of digital content is not straightforward” in that it “requires action and intervention throughout the lifecycle, far earlier and more frequently than” with physical collections. The strategy, which was approved in March 2013, outlines four main strategic priorities [British Library 2013b], i.e. to:

- *Ensure [the Library’s] digital repository can store and preserve […] collections for the long term;*
- *Manage the risks and challenges associated with digital preservation throughout the digital collection content lifecycle;*

- *Embed digital sustainability as an organisational principle for digital library planning and development;*
- *Benefit from collaboration with other national and international institutions on digital preservation initiatives.*

At least three of these priorities depend upon there being adequate knowledge of the British Library's digital collections, e.g. for being able to establish and invoke suitable preservation plans, for monitoring the wider technical environment (preservation watch), or for building awareness of digital preservation issues amongst Library colleagues and (ultimately) its users. A useful first step appeared to be to work with curators and other content specialists to develop descriptive profiles of the Library's high-level digital collection areas, with the aim of capturing key knowledge about the collections and their specific preservation requirements.

The British Library's Digital Preservation Team has, for the very first time therefore, begun to develop content profiles for the Library's high-level digital collection types. It is intended that these will help provide the opportunity to build conversations with curators and content specialists on identifying specific preservation requirements. This has a number of benefits:

- The massive scale of content held by the British Library means that collection profiling is a crucial part of preservation planning, supporting the identification of preservation requirements, and the tools necessary to facilitate these.
- Collection profiling opens a forum on which collection stakeholders, the people who make decisions at different lifecycle stages, can express challenges faced by specific content types. This should help the development of a shared understanding of digital preservation requirements from both curatorial and technical perspectives.
- Corporate understanding of the collections held by the British Library is enriched through the sharing of collection information, between the departments which make collection decisions. This acts as a platform on which to build sustainable preservation development.

3. RELATED WORK

The British Library's collection profiles are intended to support the planning of digital preservation activities across different content lifecycles. It, therefore, builds on previous work focused on the assessment of content, including the use of decision support tools for preservation planning, the use of tools and registries for content profiling or characterization, as well as more direct attempts to capture curatorial 'intent' for specific collections. There is also a link to institution-level assessment (e.g. repository audit) in that audit tools and maturity models could potentially also be applied at collection or ingest work stream level. The work is also related to ongoing research on defining the significant properties (or characteristics) of digital objects, not least in taking account of how significance may be understood differently by the various stakeholders involved in the preservation process, including creators, custodians and consumers [Knight and Pennock 2009; Dappert and Farquhar 2009].

This section will outline some related work on the assessment of collections for digital preservation, focusing on preservation planning decision-support tools (e.g. Plato), technical content characterization tools, the capture of preservation intent, and assessments at the institution or repository level.

3.1 Preservation planning

The OAIS Model defines a Preservation Planning Functional Entity that "provides the services and functions for monitoring the environment of the OAIS, providing recommendations and preservation plans to ensure that the information stored in the OAIS remains accessible to, and understandable by, the Designated Community over the Long Term, even if the original computing environment becomes obsolete" [ISO 14721:2012; CCSDS 650.0-M-2 2012]. It also provides some specific examples of what functions might be required:

Preservation Planning functions include evaluating the contents of the Archive and periodically recommending archival information updates, recommending the migration of current Archive holdings, developing recommendations for Archive standards and policies, providing periodic risk analysis reports, and monitoring changes in the technology environment and in the Designated Community's service requirements and Knowledge Base. […] Preservation Planning also develops detailed Migration plans, software prototypes and test plans to enable implementation of Administration migration goals.

It is clear from this that preservation planning is a critical component of any digital preservation strategy.

One attempt to develop a structured approach to preservation planning is the Plato decision-support tool developed as part of the Planets (Preservation and Long-term Access through Networked Services) project [Becker et al 2008; Becker et al 2009]. Plato provides a methodology and a software tool to support the systematic capture of preservation requirements from various stakeholders and then to match these to potential preservation strategies for further analysis. The result is a recommendation that can form the basis of a preservation plan, which contains information on contexts as well as the evidence base underpinning the decision.

There have been various attempts made to integrate Plato with other digital preservation systems. For example, researchers from the KeepIt and Planets projects integrated Plato and other digital preservation tools with the ePrints repository software, creating plugins to ePrints that would support the development of preservation workflows, including the generation of preservation plans and action plans [Hitchcock et al 2010].

The SCAPE (Scalable Preservation Environments) project[1] has also been exploring how to integrate Plato with other digital preservation tools and services [May and Wilson 2014]. The project is specifically interested in enabling Plato to:

- Import information from external sources, e.g. from content profiles or from institutional policies.
- Integrate with other services, e.g. the SCAPE's Component Catalogue of tools or the Scout automated preservation watch service [Faria 2013]
- Incorporate planning functionality within repository systems, so that plans can be fed back for monitoring

In terms of SCAPE, the resulting Preservation Plans document collections, their institutional context, and the decision-making process that led to the selection of a particular preservation action. It also contains a Preservation Action Plan that contains all of the

[1] SCAPE project. Retrieved August 30, 2014 from http://www.scape-project.eu/

information necessary to apply the preservation action as well as an Executable Action Plan that can be deployed through a workflow management system (e.g. Taverna).

3.2 Content characterization

Preservation planning support tools like Plato depend upon there being accurate information about the file representation types (e.g. formats) present in a collection or repository. The scope of this has been outlined by Faria et al [2013]

Digital preservation starts by understanding what content a repository holds and what are the specific characteristics of that content. This process is supported by the characterization of content and allows a content owner to be aware of content volumes, characteristics, format distributions, and specific peculiarities such as digital rights management issues, complex content elements, or other preservation risks.

Several different tools and services have been developed to help with content identification, validation, and characterization, of which JHOVE (JSTOR/Harvard Object Validation Environment) and DROID (Digital Record Object Identification) are perhaps the most well-known. Characterization software like JHOVE, JHOVE2 or DROID can in turn be embedded into other tools. For example, the File Information Tool Set (FITS), originally created by Harvard University Library, combines a number of different open source tools – currently including JHOVE, DROID, Apache Tika, and the National Library of New Zealand Metadata Extractor – in order to support consistency of use across all tools and to produce standardized output metadata [McEwen and Goethals 2009].

It is obvious that with ever growing collections, characterization tools need to work at scale. One recent initiative has been c3po (Clever, Crafty, Content Profiling of Objects), which has produced a prototype software tool that produces content profiles of collections based on data generated by FITS that can be used for further analysis or visualization [Petrov and Becker 2012]. Tools like DROID and Apache Tika have also been used to analyze very large collections, e.g. Web archives, where there is considerable interest in the use of scalable characterization tools [Jackson 2012; Palmer 2014].

The British Library has an active interest in content characterization tools, not least through its involvement in the SCAPE project, one of whose objectives is enabling the large-scale characterization of digital objects [Van der Knijff 2011]. The British Library Digital Preservation Team's current work-plan also contains work-streams for file format assessment; tool assessment and preservation watch, all of which will involve some level of content characterization at a technical level.

3.3 Content profiling

While the technical aspects of content characterization remain important, the British Library's collection profiling activity described in this paper has primarily drawn its inspiration from other content profiling activities, i.e. those based on a structured dialogue with curators and other content specialists. As part of the collection profile development, a number of content-based profile initiatives were reviewed, in particular the Digital Content Reviews (DCR) for Life Cycle Management developed by MIT Libraries and the Data Curation Profiles developed by Purdue University Libraries.

Purdue's Data Curation Profiles are a tool for capturing basic information about research datasets in order to support their curation and reuse. The profile provides a framework (an interview structure) that can be used to gather information about datasets and their potential re-use. Once completed, profiles can help guide decision-making about the management of datasets as well as inform those providing research data management services of any specific requirements [Witt et al 2009]. The Data Curation Profile toolkit (an interviewers' manual/worksheet and user guide) has been made freely available, and the profiles have begun to be used in other initiatives, e.g. by Cornell University Library to help design the Datastar research data registry [Wright et al 2013]. While the Data Curation Profiles were probably too focused on one particular type of content to be useful for our immediate purposes, the general approach clearly demonstrated the benefits of using content profiles to support lifecycle management.

MIT Libraries' Digital Content Reviews for Life Cycle Management took a similar lifecycle-management view, but – more like the emerging British Library profiles - were primarily intended to help capture information about the implications of collecting certain types of digital content. The section headings are a mixture of generic (content overview, collection management, rights management) and those that follow the content lifecycle (acquisition, ingest, preservation planning, archival storage, long-term access) [MIT Libraries 2013].

3.4 Preservation intent

While these existing content profiles provided us with a basis for developing a draft framework for the British Library profile, another key inspiration was the National Library of Australia's work on identifying 'preservation intent' [Webb et al 2013]. As part of their approach to preservation planning, digital preservation specialists at the National Library of Australia have been concerned to talk to content specialists (collection managers, curators) in order to develop some 'plain-language' statements about "which collection materials, and which copies of collection materials, need to remain accessible for an extended period, and which ones can be discarded when no longer in use or when access to them becomes troublesome." Content specialists were also "asked to make broad statements clarifying what 'accessible' means by stating the priority elements that need to be re-presented in any future access for each kind of digital object type in their collections." This both becomes a means of ensuring that curators and other collection specialists take responsibility for deciding what will happen to collections and is essential for preservation planning. Webb et al [2013] write that "without it, we are left floundering between assumptions that every characteristic of every digital item has to be maintained forever (almost certainly an impossible expectation) and assumptions that it is good enough to store data safely and let future users worry about how to access it (almost certainly an inadequate response)." Capturing elements of preservation intent seemed vital for the success of the British Library's collection profiling activity.

3.5 Institution-level assessment

Other approaches to digital preservation assessment have been focused on higher levels of aggregation than collections. This includes well-established work on repository audit, where the main focus of attention has been on two interrelated standards:

- The Trustworthy Repositories Audit & Certification (TRAC) criteria and checklist published by the US Center for Research Libraries [2007]
- ISO 16363:2012 Audit and Certification of Trustworthy Digital Repositories [ISO 16363:2012].

Both provide a framework for the assessment of repositories based on three main categories: organizational infrastructure (including governance, structure and financial sustainability), digital object management, and infrastructure and security risk management.

These standards mainly focus on organization and infrastructure rather than collections, but some other approaches to institutional evaluation do have the potential to be able to inform the assessment at collection-level. This is particularly true of approaches based on maturity modelling, which include the Digital Preservation Capability Maturity Model, whose levels mainly focus on perceived risks to content, but whose assessment categories specifically take into account things like policies, governance and expertise, i.e. taking into account significant organizational and human factors [Dollar and Ashley 2013]. The role of maturity models is also being actively explored in the research data management domain, both at organization and community levels [Crowston and Qin 2010; Lyon et al 2012].

A similar approach has been taken by the US National Digital Stewardship Alliance in developing the NDSA Levels of Digital Preservation, which are understood to be "a tiered set of recommendations on how organizations should begin to build or enhance their digital preservation activities" [Phillips et al 2013]. The NDSA Levels provide technical guidance on preserving digital content "at four progressive levels of sophistication across five different functional areas," which are:

- Storage and geographic location
- File fixity and data integrity
- Information security
- Metadata
- File formats

The NDSA Levels are deliberately focused on the technical aspects of digital preservation as the team wanted them "to focus on practices, not policies or workflows, in order to allow immediate implementation" [Phillips et al 2013]. As it turns out, the functional areas identified by the NDSA correspond quite well to the types of information required for assessment at collection level.

4. COLLECTION PROFILING

The development of collection profiles at the British Library has been broken down into a number of smaller steps. The initial tasks were to identify the British Library's high-level digital collection areas and to develop an initial template to capture the required information [Day et al 2014].

4.1 Identifying high-level collection types

An initial practical task was to identify and define what we understood to be the Library's high-level digital collections. There was no agreed list of digital collection types held by the Library. Those lists that did exist - e.g. those provided by the catalogue or website - often included, for reasons of practicality, content types at several different levels of granularity.

In order to arrive at a more consistent list of candidate collection types, it was decided to supplement the categories found in these ad hoc lists with others derived from the Library's digital asset register. It is important to recognize that we were not trying to produce a definitive taxonomy of all digital collection types held by the Library, but simply to be able to identify collections at a sufficient (and logical) level of granularity in order to get started on the development of content profiles. The high-level collection types eventually identified (Table 1) included some that were firmly based on resource type (e.g. sound, multimedia), others that were multi-faceted but based on particular content streams (e.g. web archives); and others that followed more traditional categorizations of library collections, updated for the digital era (e.g. journals, books).

Table 1. Initial High-Level Collection Types

Type	Collection
Newspapers / journals	Digitised newspapers
	Born digital newspapers
Books	NDLP eBooks
	Voluntary deposit
	Digitised printed books
	Turning the Pages content
Manuscripts / Archives	Digitised Manuscripts
	Digitised archives
	Personal digital archives
	Turning the Pages content
Music	Digitised Music Collections
	Sheet Music
Maps	Digital mapping supplied by Ordnance Survey (GIS)
	Digitised maps
Academic journals	NPLD eJournals
	Voluntary deposit e-Journals
	Subscription e-Journals
Theses	Digitised theses
Patents	Patent databases
Web archives	UK Web Archive
	NDLP Web domain harvests
Sound / multimedia	Archive sound recordings
	Sound Archive (e.g., field recordings)
	Digitised sound / video
Stamps	Digitised stamps
Photographs	Digitised photographs
Printed ephemera	Digitised ephemera

The process of developing a list of high-level collection types, however, did raise some interesting questions about the task we had set ourselves.

4.1.1 Born-digital vs digitized content

For example, it was initially tempting to categorize digitized content separately from 'born-digital,' as this is a familiar distinction made by those considering digital preservation [Daigle 2013]. However, part of the aim of the profiling work was to try to deal with content by type, regardless of provenance or format.

So, for example, the British Library's digital newspaper collections would potentially include:

- Digitised printed newspapers from the Library's own collections (e.g. the historical newspaper collections digitized in collaboration with Gale Cengage, typically comprising images with searchable OCR text)
- E-editions of printed newspapers, ingested directly from newspapers' publication workflows (e.g. as PDF)
- Web-based newspapers captured as Web archives (e.g. newspaper websites captured as part of the UK Web Domain; the originals are typically constantly evolving Web pages with significant amounts of embedded content (e.g. images, video, surveys, comments) and links)

Obviously, within the Library's ingest and processing workflows these would be represented by quite different content streams, but the profiling activity does at least give an initial opportunity to consider all digital news content as a single collection, even if it is decided later on that more than one kind of preservation intent can be identified. Similar considerations would apply to other kinds of content.

At a more fundamental level, however, it is increasingly difficult to distinguish between born-digital and digitized content. As others have pointed out, much digital content is often a combination of several different kinds of content type, some of which may be born-digital, others not [Friedlander 2002]. This is perhaps most noticeable with Websites, but is increasingly true of many other kinds of content, For Example, e-journal articles or e-books could be understood to be simply containers for multiple kinds of content, which might include images, video, sound, games, software or data. In Europe, at least, research papers reimagined as compound digital objects (combining at least text and data) are sometimes known as "enhanced publications" [Doorenbosch et al 2009]. Eventually, as predicted by Kircz [1998], it might also be possible to think of all research papers as modular aggregations of many other kinds of content, including bibliographic information, content, abstracts, references, index terms, tables, etc., all of which could potentially have a different representation.

All of this meant that we needed - at least to start with - to focus on content type regardless of its immediate provenance.

4.1.2 The 'tangibility' of collections

When developing the collection profile activity, we also had to understand what exactly we meant when we talked about "collections"? Collections are a deeply embedded concept in memory institutions, so quite a lot of intellectual effort has been made over the years into trying to understand what they are and how they relate to wider organizational contexts. Traditional concepts of collection in library and information science have tended to focus on three main things: tangibility (regardless of format), ownership and a perceived user community [Lee 2000]. What has changed in the digital era is that library collections can be built without the inherent need for tangibility (although even digital content has to be stored somewhere) or ownership.

Like most other research libraries, the British Library routinely provides access to digital content that is not under its own direct control. As mentioned before, its current content strategy states that outside of legal deposit, voluntary deposit and donation, "the Library will prefer to connect to content," unless it wishes "to hold and preserve the materials for the long term" [British Library 2013a]. In this new collection management environment, active choices need to be made about precisely which content needs to become part of the permanent collections (and is thus able to be preserved). It is intended that the collections profiling activity at the Library will support collections management decision making, not least by gaining insight from collection specialists and curators on the specific preservation requirements of different classes of content. It might also help to clarify which particular content needs to become part of the Library's permanent collections.

Table 2. Initial Profile Framework Structure

Summary	Content Type (from list).
	Brief Description.
	Location.
	Curators / collection owners.
	Interviews held.
	Legal Deposit status.
	Creation status.
	Accrual status.
	Number of digital objects (approximate).
Background	An introduction to the content type, providing background on the collection/s covered by the profile.
Acquisition	Identifying the main current acquisition routes for collection content.
Preservation Intent	Summary of points agreed by curators / content owners, identifying the main characteristics of collections that will need to be preserved.
Acquisition Format	Identifying the main formats currently being acquired (where collections are complex, this does not need to be exhaustive).
Issues	Highlighting any specific current challenges.
Profile Metadata	Information about the completed collection profile itself, e.g. identifying creators, dates, and status / version number.

5. Developing the draft profile framework

The framework for the profile itself was developed at the same time as the identification of high-level collection types. The sections in the initial draft profile framework (November 2013) section headings were either generic (collection overview, preservation intent, rights) or broadly followed the functions defined by the Reference Model for an Open Archival Information System (ingest, archival storage, preservation planning, access control). Following the review of some draft profiles, the framework has been further simplified to reduce the number of sections required and to focus the profile on the key information types required to support the capture of digital preservation requirements (Table 2).

The draft framework was first introduced to and discussed with curatorial and other colleagues in the Library. It was then used to help create a number of draft profiles, initially for content types covered by Non-Print Legal Deposit content streams (e-journals, e-books, UK Web-domain harvests), then by a few selected others (manuscripts and archives, news content, sound content).

The profile framework will evolve further as we gain more experience with using it. Eventually, however, the plan will be to develop some support materials (e.g. documentation, a set of sample interview questions) that will help with ensuring consistency of approach. It will also be important to review the profiling process following integration with other preservation planning activities being undertaken by the British Library (e.g. file format assessments, tool assessments and preservation watch).

It is highly likely that both collections and preservation intent will change over time. There will be a need to ensure that collection profiling is undertaken on a regular basis and that it becomes part of the Library's business-as-usual digital preservation activities.

6. DISCUSSION AND OUTLOOK

The British Library's collection profile activity is an attempt to use content reviewing to capture information about collections and preservation intent to help inform digital preservation planning. Work to date has included an attempt to identify the high-level digital collections in the Library and to define an initial profile framework. Work on developing the profiles is ongoing as we progress in an iterative fashion. It promises to be an interesting approach, linking curators understanding of digital collections with the planning processes required to support their digital preservation.

7. ACKNOWLEDGMENTS

The authors would like to thank British Library colleagues for their comments on the draft list of high-level collection types and the initial collection profile framework.

Ann MacDonald was Intern – Digital Preservation at the British Library from October 2013 to March 2014 and was a key part of the team developing collection profiles.

8. REFERENCES

APARSEN. 2013. The British Library. In *Storage solutions summary of inputs*. APARSEN Deliverable D23.1, 26-33. (March 2013). Retrieved August 30, 2014 from http://www.alliancepermanentaccess.org/wp-content/uploads/downloads/2013/03/APARSEN-REP-D23_1-01-1_0.pdf

Christoph Becker, Hannes Kulovits, Andreas Rauber, and Hans Hofman. 2008. Plato: a service oriented decision support system for preservation planning. In *Proceedings of the 8th ACM/IEEE-CS joint conference on Digital libraries* (Pittsburgh, PA, USA, June 16 – 20, 2008). ACM Press, New York, NY, 367-370. DOI:http://dx.doi.org/10.1145/1378889.1378954

Christoph Becker, Hannes Kulovits, Mark Guttenbrunner, Stephan Strodl, Andreas Rauber and Hans Hofman. 2009. Systematic planning for digital preservation: evaluating potential strategies and building preservation plans. *International Journal on Digital Libraries* 10 (2009), 133-157. DOI:http://doi.org/10.1007/s00799-009-0057-1

Caroline Brazier. 2013. born.digital@british.library: the opportunities and challenges of implementing a digital collection development strategy. In *Proceedings of the IFLA World Library and Information Congress* (Singapore, August 17 - 23, 2013). Retrieved August 30, 2014 from http://library.ifla.org/222/1/198-brazier-en.pdf

British Library Act 1972. Her Majesty's Stationery Office, London (1972). Retrieved August 30, 2014 from http://www.legislation.gov.uk/ukpga/1972/54/contents

British Library. 2013a. *From stored knowledge to smart knowledge: the British Library's Content Strategy 2013-2015*. British Library, London. Retrieved August 30, 2014 from http://www.bl.uk/aboutus/stratpolprog/contstrat/british_library_content_strategy_2013.pdf

British Library. 2013b. *Digital Preservation Strategy, 2013-2016*. British Library, London. Retrieved August 30, 2014 from http://www.bl.uk/aboutus/stratpolprog/collectioncare/discovermore/digitalpreservation/strategy/BL_DigitalPreservationStrategy_2013-16-external.pdf

CCSDS 650.0-M-2. 2012. Reference Model for an Open Archival Information System (OAIS). CCSDS Recommended Practice (June 2012).

Center for Research Libraries. 2007. *Trustworthy Repositories Audit & Certification: criteria and checklist (TRAC)*, v 1.0 (February 2007). Retrieved August 30, 2014 from http://www.crl.edu/archiving-preservation/digital-archives/metrics-assessing-and-certifying-0

Sheila Corrall. 2011. The concept of collection development in the digital world. In Maggie Fieldhouse and Audrey Marshall (Eds.). *Collection development in the digital age*. Facet Publishing, London, 3-25.

Sheila Corrall and Angharad Roberts. 2012 Information resource development and "collection" in the digital age: conceptual frameworks and new definitions for the network world. In *Libraries in the Digital Age (LIDA) Proceedings*, Vol. 12 (Zadar, Croatia, June 18 – 22, 2012). Retrieved August 30, 2014 from http://ozk.unizd.hr/proceedings/index.php/lida/article/view/62/33

Kevin Crowston and Jian Qin. 2010. A Capability Maturity Model for scientific data management. In *Proceedings of the 73rd ASIS&T Annual Meeting on Navigating Streams in an Information Ecosystem* (Pittsburgh, PA, USA, Oct. 22 – 27, 2010). ACM Press, New York, NY, Article 124.

Bradley J. Daigle. 2013. Dream the impossible dream: born digital stewardship. In *Archiving 2013 Final Program and Proceedings* (Washington, DC, USA, April 2-5 April, 2013). Society for Imaging Science and Technology, Springfield, VT, 2-5.

Angela Dappert and Adam Farquhar. 2009. Significance is in the eye of the beholder. In *Proceedings of the 13th European Conference on Digital Libraries* (Corfu, Greece, September 27 - October 2, 2009). Springer, Berlin, 297-308. DOI:http://dx.doi.org/10.1007/978-3-642-04346-8_29.

Michael Day, Ann MacDonald, Akiko Kimura and Maureen Pennock. 2014. Implementing digital preservation strategy:

developing content collection profiles at the British Library. In *Proceedings of Digital Libraries 2014* (London, UK, September 8 - 12, 2014). Forthcoming.

Charles M. Dollar and Lori J. Ashley. 2013. Assessing digital preservation capability using a maturity model process improvement approach. (February 2013). Retrieved August 30, 2014 from http://www.savingthedigitalworld.com/papers-research

Paul Doorenbosch, Eugène Dürr, Barbara Sierman, Jens Ludwig and Birgit Schmidt. 2009. Long-term preservation of enhanced publications. In Marjan Vernooy-Gerritsen (Ed.). *Enhanced publications: linking publications and research data in digital repositories*. Amsterdam University Press, Amsterdam, 157 -209. DOI:http://doi.org/10.5117/9789089641885

Luís Faria. 2013. Scout – a preservation watch system. In: *Open Planets Foundation blog*. (16 December 2013). Retrieved August 30, 2014 from http://www.openplanetsfoundation.org/blogs/2013-12-16-scout-preservation-watch-system

Luís Faria, Christoph Becker, Kresimir Duretec, Miguel Ferreira1 and José Carlos Ramalho. 2013. Supporting the preservation lifecycle in repositories. In *Proceedings of Open Repositories 2013* (Charlottetown, PEI, Canada, July 8 – 12, 2013). Retrieved August 30, 2014 from http://or2013.net/sites/or2013.net/files/PW_repositories_OR13_V0.5.pdf

Patrick Fleming. 2011. The British Library Newspaper Strategy: developing collaboration with publishers to digitise back runs and to ingest born digital newspapers. In Hartmut Walravens (Ed.). *Newspapers: legal deposit and research in the digital era*. IFLA Publications, Vol. 150. De Gruyter Saur, Munich, 21-30.

Any Friedlander. 2002. Summary of findings. In *Building a national strategy for digital preservation: issues in digital media archiving*. Council on Library and Information Resources, Library of Congress, Washington, DC, 1-8. Retrieved August 30, 2014 from http://www.clir.org/pubs/abstract//reports/pub106

Richard Gibby and Caroline Brazier. 2012. Observations on the development of non-print legal deposit in the UK. *Library Review* 61, 5 (2012), 362-377. DOI:http://doi.org/10.1108/00242531211280487.

Steve Hitchcock, David Tarrant, Les Carr, Hannes Kulovits and Andreas Rauber. 2010. Connecting preservation planning and Plato with digital repository interfaces. In *Proceedings of iPRES 2010, 7th International Conference on Preservation of Digital Objects* (Vienna, Austria, 19 - 24 Sep 2010). Retrieved August 30, 2014 from http://www.ifs.tuwien.ac.at/dp/ipres2010/papers/tarrant-65.pdf

ISO 14721:2012. Space data and information transfer systems -- Open archival information system (OAIS) -- Reference model. International Organization for Standardization, Geneva.

ISO 16363:2012. Space data and information transfer systems -- Audit and certification of trustworthy digital repositories. International Organization for Standardization, Geneva.

Jackson, A. 2012. Formats over time: exploring UK Web history. In *Proceedings of iPres 2012* (Toronto, October 1 - 5, 2012). University of Toronto, Faculty of Information, Toronto, Ontario, Canada, 155-158. Retrieved August 30, 2014 from https://ipres.ischool.utoronto.ca/proceedings

Joost G. Kircz. 1998. Modularity: the next form of scientific information representation? *Journal of Documentation* 54, 2 (1998), 210-235. DOI:http://doi.org/10.1108/EUM0000000007185

Gareth Knight and Maureen Pennock. 2009. Data without meaning: establishing the significant properties of digital research. *International Journal of Digital Curation* 4, 1 (2009), 159-174. DOI:http://dx.doi.org/10.2218/ijdc.v4i1.86

Hur-Li Lee. 2000. What is a collection? *Journal of the American Society for Information Science* 51, 12 (2000), 1106-1113. DOI:http://doi.org/10.1002/1097-4571(2000)9999:9999<::AID-ASI1018>3.0.CO;2-T

Legal Deposit Libraries (Non-Print Works) Regulations 2013. S.I. 2013 No. 777. The Stationery Office, London (2013). Retrieved August 30, 2014 from http://www.legislation.gov.uk/uksi/2013/777/contents/made

Liz Lyon, Alex Ball, Monica Duke and Michael Day. 2012. Developing a Community Capability Model Framework for data-intensive research. In *Proceedings of iPres 2012* (Toronto, Ontario, Canada, October 1 - 5, 2012). University of Toronto, Faculty of Information, Toronto, Ontario, 9-16. Retrieved August 30, 2014 from https://ipres.ischool.utoronto.ca/proceedings

Peter May and Carl Wilson. 2014. *Technical architecture report*, v2. SCAPE Deliverable D2.3 (March 2014). Retrieved August 30, 2014 from http://www.scape-project.eu/deliverable/d2-3-technical-architecture-report-v2

Spencer McEwen and Andrea Goethals. 2009. File Information Tool Set (FITS): a new tool for digital preservation repositories. *D-Lib Magazine* 15, 9/10 (September/October 2009). Retrieved August 30, 2014 from http://www.dlib.org/dlib/september09/09inbrief.html

MIT Libraries. 2013. Digital Content Management Infrastructure Improvement: FY13 strategic objective. Retrieved August 30, 2014 from http://libguides.mit.edu/lifecycle

William Palmer. 2014. A Tika to ride: characterising web content with Nanite. In *Open Planets Foundation blog*. (21 March 2014). Retrieved August 30, 2014 from http://www.openplanetsfoundation.org/blogs/2014-03-21-tika-ride-characterising-web-content-nanite

Petar Petrov and Christoph Becker. 2012. Large-scale content profiling for preservation analysis. Poster presentation, iPres 2012 (Toronto, Ontario, Canada, October 1 - 5, 2012). Retrieved August 30, 2014 from http://ifs.tuwien.ac.at/~petrov/publications/c3po-poster-ipres12.pdf

Megan Phillips, Jefferson Bailey, Andrea Goethals and Trevor Owens. 2013. *The NDSA Levels of Digital Preservation: an explanation and uses*. National Digital Stewardship Alliance, Washington, DC. Retrieved August 30, 2014 from http://www.digitalpreservation.gov/ndsa/working_groups/documents/NDSA_Levels_Archiving_2013.pdf

Johan Van der Knijff and Carl Wilson. 2011. *Evaluation of characterisation tools, part 1, identification*. SCAPE Technical Report (September 2011). Retrieved August 30, 2014 from http://www.openplanetsfoundation.org/system/files/SCAPE_PC_WP1_identification21092011.pdf

Colin Webb, David Pearson and Paul Koerbin. 2013. 'Oh, you wanted us to preserve that?!' Statements of preservation intent for the National Library of Australia's digital collections. *D-Lib Magazine* 19.1/2 (Jan./Feb. 2012). DOI:http://dx.doi.org/10.1045/january2013-webb

Peter Webster. 2013. Crawling the UK web domain. In *UK Web Archive blog*. (16 September 2013). Retrieved August 30, 2014 from: http://britishlibrary.typepad.co.uk/webarchive/2013/09/domaincrawl.html

Michael Witt, Jacob Carlson, D. Scott Brandt and Melissa H. Cragin. 2009. Constructing Data Curation Profiles. *International Journal of Digital Curation* 4, 3 (2009), 93-103. DOI:http://doi.org/10.2218/ijdc.v4i3.117

Sarah J. Wright, Wendy A. Kozlowski, Dianne Dietrich, Huda J. Khan, Gail S. Steinhart and Leslie McIntosh. 2013. Using Data Curation Profiles to design the Datastar dataset registry. *D-Lib Magazine* 19, 7/8 9 (July/August 2013). DOI:http://dx.doi.org/10.1045/july2013-wright

DRM and digital preservation: A use case at the German National Library

Stefan Hein
Deutsche Nationalbibliothek
Adickesallee 1
60322 Frankfurt am Main, Germany
+49 69 1525 1722
s.hein@dnb.de

Tobias Steinke
Deutsche Nationalbibliothek
Adickesallee 1
60322 Frankfurt am Main, Germany
+49 69 1525 1762
t.steinke@dnb.de

ABSTRACT

Digital Rights Management (DRM) is in use for many digital publications. Digital libraries with a mandate to collect and preserve publications have to deal with technical challenges for preservation of DRM restricted objects. In the European project APARSEN a systematical classification of DRM methods and its risks for digital preservation was introduced. The German National Library handles the different types of DRM protections within the ingest workflow of the archival system by analysis and case-by-case distinction.

General Terms

Management, Measurement, Standardization, Verification.

Keywords

Digital rights management, Digital preservation, Deutsche Nationalbibliothek, APARSEN, Ingest Level

1. INTRODUCTION

Digital Rights Management (DRM) of digital publications like e-books, multimedia disks and audio files could come in many different ways. Restrictions for access and usage are often implemented by technical measures. Typical restrictions are prevention of creating copies or usage in not allowed software environments.

Libraries collect, archive and give access to digital publications. The challenge of digital preservation is dealing with obsolescence of hardware and software. Especially for national libraries with a mandate to preserve their collected objects for an unlimited time, dedicated preservation strategies and actions are needed. File format migration and emulation of old systems environments are common ways to handle the task.

The technical measures of DRM could be a problem for preservation actions. File format migration means converting and copying files, emulation means using an object in another technical environment. Both strategies might be in opposite to the intended restrictions of DRM. There are also other potential problems like dependencies on online sources for verification.

The German National Library, Deutsche Nationalbibliothek (DNB), collects many different types of digital publications within legal deposit legislation. As a partner of the European project APARSEN[1] DNB worked on a systematical approach to classify the challenges that DRM could be for digital preservation.

2. DRM: A CHALLENGE FOR DIGITAL PRESERVATION

Through the integration of proprietary rights control mechanisms as an integral component of digital objects, a new problem has arisen regarding long-term preservation (LTP). The main cause of this problem has been that restrictions of access and usage could hinder the preservation of the object. If access to the content is already blocked, the problems involved in executing LTP measures are clearly apparent. Preservation measures without access to the actual content are not viable. Technical or other types of metadata (e.g. bibliographic) can only – if at all – be extracted to a limited extent from protected files. According to OAIS, however, these data need to be incorporated in the data management and are essential for meaningful preservation planning and the execution of preservation actions ([1]). The encrypted content could also conceal malware (viruses, trojans) which could enter the archive and remain undiscovered by virus scanners.

2.1 Scale for Long-Term Preservation Risk

In order to evaluate the risk of different DRM technologies, APARSEN defined the following scale (Long-Term-Preservation Risk (LPTR)):

Table 1. Long-Term-Preservation Risk (LTPR)

LTPR	Characterization
no risk	No risk for future LTP measures
medium	Possible to use at present (at time of publication) in up-to-date hardware and software environment, current LTP measures restricted, no external dependencies, medium risk for future LTP measures
high	Use and LTP measures already (currently) restricted, high risk for implementation of LTP measures in the future as result of external dependencies

In summary, the higher the LTPR value, the greater the risk in archiving and maintaining the usability of the object concerned.

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

This appraisal contains a prediction component, meaning that 100% guarantees cannot be offered.

2.2 Classification and Assessment

In the APARSEN “Report on DRM preservation” ([2]) four DRM variants are identified and assessed:

Data carrier copy protection, LTPR = medium: Data carrier migration is a key LTP measure, meaning that the prevention of all activities aimed at separating the data stream from the carrier should be regarded as risky. The data carrier copy protection prevents in principle copying. If the data stream cannot be separated from the data carrier, this carries a high risk for future LTP measures because the necessary players and/or software may no longer be available. Usage is, however, possible at present with common player devices. Depending on the kind of data carrier protection, data carrier migration might be possible with current equipment, albeit with restrictions e.g. a loss of quality in case of a digital-analogue conversion of audio material.

Lightweight DRM, LTPR = no risk: Lightweight DRM (LWDRM) refers to all mechanisms which do not of themselves restrict access to digital objects or their use, but which serve the detection and tracking of legal infringements [3]. This is mostly achieved through the use of marking techniques such as digital watermarks. Digital watermarks may be applied to the digital object in a way which is invisible to the user but which allows the content providers to detect their works e.g. on illegal file-sharing sites. Lightweight DRM involves no restrictions on access or usage. The marking of digital objects therefore poses no risk for use or LTP measures.

Encryption-based password protection, LTPR = medium: This variant focuses on DRM mechanisms which require no connections to external components (such as authentication servers) during use and which basically manage the access and usage possibilities of objects. The term "access" here signifies the opening of a file object using pre-defined player and display software - even though the act of opening could itself be interpreted as the most basic form of use. Use is therefore always conditional upon having access to the object. An example of this is Adobe's PDF format. It contains functions which render access and usage and it is manageable in a variety of forms (like print, edit document, copy content, extract pages). This kind of limitation of use is one of the most common DRM variants that libraries such as the German National Library face, primarily in the context of online publications (e.g. e-books) and dissertations. The access to the data stream and the use of the content is predicated upon knowing the password. The password must be saved separately and linked to the actual content. The user must be given the password when access is granted. If only limited usage rights, such as text extraction, are granted yet the content can still be displayed, it can no longer be predicted with any certainty whether the conversion tool will require precisely this feature in the future. The execution of current and future LTP measures therefore carries risks.

DRM Systems, LTPR = high: This DRM category focuses not only on selected aspects already presented above, but also attempts, by means of a system of diverse components and technologies such as the digital watermarks and encryption methods already examined, to cover all the core DRM areas. The architecture of a DRM system (figure 1) is outlined by Bill Rosenblatt ([4]) and consists of the three linked components of content server, licence server and client. The different DRM components can be geographically distributed and communicate via the Internet. This results in a range of dependencies which can affect everything from generation and content through to use. The client, e.g. the media player or the document reader, therefore no longer functions independently as a gateway to the actual content. It is apparent that precisely this interaction between the different components markedly increases the complexity of DRM systems in comparison to the DRM variants already presented.

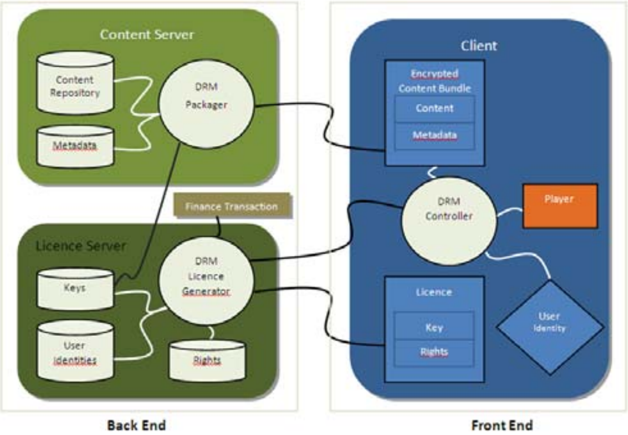


Figure 1, Architecture of a DRM System (adapted from [4])

Given that access to and use of the content is restricted similar to the "encryption-based password protection” variant, objects protected by DRM systems also carry the same risks. A further problem factor is the existence of an external license server, and connection to it is a precondition for encryption. Even today, use may be impaired or prevented entirely in the event of the content provider going out of business, network problems etc.

3. DRM AT THE GERMAN NATIONAL LIBRARY

DNB takes care that all digital publications can be utilized in accordance with legal regulations. Depending on the rights that the content producer grants DNB during the submission process, some publications can be provided in-house only, while others are remotely accessible. DNB receives DRM protected material but does not produce material that is DRM protected. In general publications which are published by the DNB are DRM free. Also DNB advises its deliverers to abstain from the use of DRM mechanism for the delivery to the DNB. In the past DRM mechanism of digital objects were only detected manually. However, no statistical recordings of DRM mechanisms detected were implemented. But it can be assumed that the proportion of DRM protected material has been increasing in parallel to the further development of DRM techniques and format capabilities.

3.1 Data Types

The following data types are occasionally submitted with integrated DRM measures to the DNB:

- Doctoral theses and teaching theses of German universities
- DNB digitized print media
- e-books
- e-journals

- e-papers

The use of DRM techniques and tools depends on the file format and its capabilities, the data type and the publisher. The following techniques were detected so far:

- PDF document restrictions (password protection and print, copy restrictions)
- Adobe's LifeCycle Management (mostly publishers)
- encrypted ZIP container

3.2 Approach

DNB considers DRM measures as a potential risk to fulfill its legal obligation. Since the end of 2012, DNB uses tools to detect DRM measure of digital objects during the ingest process. Before that time the detection was manually done by random sampling.

In accordance with the decision to preserve unaltered originals and to abstain from normalization measures at the time of ingest, the DNB tries to collect the unprotected version of the digital object whenever it is possible.

The approach for online publication contains the decision to refuse "DRM suspicious" material after detection and give the publisher or the delivering institution the possibility to remove the protection for a second delivery. "DRM suspicious" means the existence of DRM techniques which were assessed as medium or high (LTPR).

The Ingest Level concept that is in use at the German National Library leads to provisional rejection of all objects with any kind of DRM ([5]). An Ingest Level is an assigned risk of preservation. This is based on five criteria: file integrity (FI), file format identification (ID), technical restrictions (TR), format specific metadata (MD) and file format validity (V). These criteria are automatically checked within the ingest workflow and an Ingest Level of 0 to 4 is assigned (table 2). Any kind of DRM restriction means level 0 or 1 and a provisional rejection.

Table 2, Ingest Level and criteria

	FI	ID	TR	MD	V
Level 0	X	O	O	O	O
Level 1	X	X	O	O	O
Level 2	X	X	X	O	O
Level 3	X	X	X	X	O
Level 4	X	X	X	X	X

It is, however, not always an option to reject DRM protected objects, respectively, to request DRM free versions, especially when the producer cannot be identified anymore. Furthermore, not every content provider is immediately willing to provide its objects without DRM to the preservation institution.

In these cases, it can only be attempted to create awareness for the problem on the side of the producer / content provider. In the case of the DNB, the legal mandate can be used as an argument. Also the guarantee that the rights will be protected via an institutional access management, so that no disadvantages result from DRM free objects for the content provider, can assist the argumentation. This approach, however, implies additional effort, namely in the implementation of such an access management.

For the automatic detection DNB uses the support of open-source tools. In the case of encrypted ZIP containers the regular unpack routine would report the protection measure. For some time now the automatic generation of technical metadata using metadata tools has been a recognized and established component of the ingest process. The DNB has long been using the File Information Tool Set (FITS)[2] as a framework for using an entire tool set. This framework provides access to a whole range of tools including the JSTOR/Harvard Object Validation Environment (JHOVE and JHOVE2)[3] tool, the Digital Record Object Identification (DROID)[4] tool and the NLNZ Metadata Extractor. Use of a tool set widens file format support and reduces the risk of errors in the identification and validation of the file format. Some of the above tools (e.g. JHOVE) also permit the recognition of document restrictions such as password-protected PDF files.

As a wrapper for FITS we use a self-developed tool called didigo (diagnose digital objects). FITS is called from didigo for every file and the FITS output of the different analysis tools is used to calculate the Ingest Level. The Ingest Level is compared to the expected value for the files and actions are initiated accordingly.

One result of the automated ingest routine, the provisional rejection of DRM protected objects and the request for re-submission of unprotected material is that the number of ingested DRM protected PDFs in the DNB collection has been very low since the end of 2012: Only 146 PDF documents out of a total of 1,630,600 PDF documents that were ingested between December 2012 and March 2014 are DRM protected (figure 2).

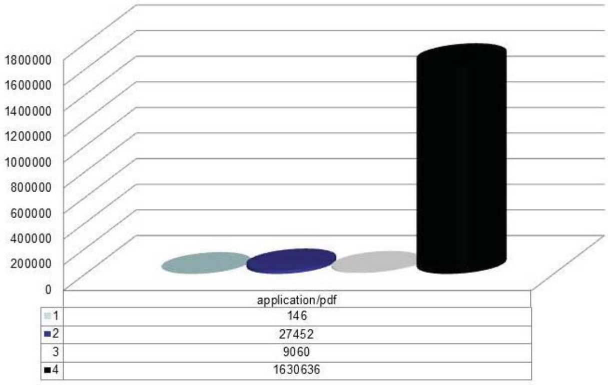


Figure 2, Number of PDF files per Ingest Level

According to its legal mandate the DNB takes preservation actions like migration on archived publications. Where DRM mechanisms inhibit preservation actions, an agreement between the German Publishers and Booksellers Association, the national association of the phonographic industry and the DNB, allows the DNB to remove DRM mechanisms for archival purposes. In particular this is important for post processing the stock of already archived objects, which have unrecognized DRM mechanisms.

As mentioned before, DRM was only detected manually in sample checks between 1998 and 2012. DNB has accepted quite a

[2] http://fitstool.org/

[3] http://jhove.sourceforge.net/

[4] http://www.nationalarchives.gov.uk/information-management/projects-and-work/droid.htm

number of DRM protected objects into the archive. This was, however, not documented, and therefore, no statistical figures are available.

During 2014, DNB will re-ingest all “old” objects in the collection with the new, automated ingest workflow that makes use of several metadata tools. With this, DNB will at least, or in a first step, be able to identify the DRM protected objects in its collection for further treatment.

Based on the statistical findings and the DRM analysis, it is possible to plan countermeasures. This will probably become a project in its own right. Where possible, DNB will try to get in touch with the publishers and request re-submission or will try to remove the DRM protection.

3.3 Limits

One limit of the approach of refusing “DRM suspicious” material lays in the limited capabilities of the used metadata tools. So the tools have to be up to date to support new formats and format versions. Unfortunately FITS is not able to determine all variants of PDF restrictions. But if that would be possible another question would arise: Which restrictions are real risks for long-term preservation activities? If the user is not allowed to print the document, it might not be a risk for a conversion in the context of format migration actions. In cases of format transformations a further question still arises as to whether and how such usage restrictions should be preserved.

The alternative approach of removing DRM mechanisms implies many problems in itself. Removing technical mechanisms needs corresponding tools and might change the authenticity of the object. In general it is not easy to acquire a software tool that violates the current legislative. If there aren’t any tools or they are not allowed to use the last approach for encrypted documents could be trying every combination of possible password characters. That approach is known as a brute-force attack and is very expensive, because it needs a lot of hardware resources like processor time. For long password lengths it takes a very long time to crack the password, in the worst case the cracking attempts are nearly infinite.

4. CONCLUSION

Technical measures of DRM can be classified in four categories. The most critical category for digital preservation is related to external dependencies like online verification. Local encryption and hardware protection might be a serious threat for preservation actions as well, but there could be ways to maintain access by specific solutions or agreements.

The German National Library uses file analysis tools within the ingest workflow to recognize and categories possible threats for digital preservation. If a protection with high or medium risk is detected the publishers are requested to re-submit the files without protection. Older collected objects with protections could be a problem. DNB has an agreement with the right holders that allows the removal of technical protection measures for archiving proposes, but this was not yet done. In future projects the existing collections will be checked and protected files will be changed if it is possible and feasible.

In general the increase and change of file formats, their implementations and the DRM techniques that they contain are some of the biggest challenges. Therefore it is necessary to keep the used analyzer tools and reading platforms up to date. Furthermore new technologies like tablet PCs and portable e-book readers with new embedded techniques to protect digital rights have to be considered.

It is important to detect DRM measure as early as possible – then there is a good chance to contact the author or publisher for a DRM-free version. The more time has passed, the smaller the chance to get in contact with the rights holder. That increases the risk to have to deal with a restricted version of a publication for preservation and access.

5. REFERENCES

[1] CCSDS, "Reference Model for an Open Archival Information System (OAIS)," June 2012. [Online]. Available: public.ccsds.org/publications/archive/650x0m2.pdf. [Accessed 26 11 2013].

[2] K. Kaur, S. Hein, S. Schrimpf, M. Ras and M. Holzmayer, “Report in DRM preservation”, 2014. [Online]. Available: http://www.alliancepermanentaccess.org/wp-content/plugins/download-monitor/download.php?id=D31.1+Report+on+DRM+preservation. [Accessed 06 03 2014].

[3] R. Grimm and C. Neubauer, "LWDRM - An Alternative Rights Management System," 2004. [Online]. Available: http://waste.informatik.hu-berlin.de/Grassmuck/drm/Folien-Grimm-Neubauer-eng.pdf. [Accessed 25 11 2013].

[4] B. Rosenblatt, "Enterprise Digital Rights Management," 14 July 2005. [Online]. Available: http://www.giantstepsmts.com/Authentica-RMS%20Whitepaper.pdf, pg. 5. [Accessed 06 03 2014].

[5] Schmitt, K., & Hein, S., “Risk Management for Digital Long-Term Preservation Services”, from IPRES 2013 : proceedings / of the 10th International Conference on Preservation of Digital Objects: http://purl.pt/24107/1/iPres2013_PDF/Risk%20Management%20for%20Digital%20Long-Term%20Preservation%20Services.pdf. [Accessed 06 03 2014].

Shaping a national consortium for digital preservation

Darryl Mead
National Library of Scotland
George IV Bridge
Edinburgh
+44 131 623 3731
d.mead@nls.uk

ABSTRACT

This paper asks the question "what form of digital preservation collective is best for Scotland?" and then sets out the options being explored under the leadership of the National Library of Scotland (NLS). As a result this paper straddles several areas across the themes of the conference. It deals with the desire for the creation of an integrated national preservation **infrastructure** in Scotland. It also looks at the ways we can develop an appetite for collaboration to align differing institutional contexts for preservation to build a better **community**. Additionally it examines issues around working within the **strategic environment** to coordinate local, regional and national approaches across Scotland and the United Kingdom.

General Terms

Infrastructure, communities, strategic environment.

Keywords

Scotland, National Library of Scotland, digital preservation coalition building.

1. INTRODUCTION

I present this paper, not as a technical expert in digital preservation, but as an industry leader looking to develop a rational national response to the practical application of digital preservation across the entire cultural heritage sector of my country.

At the same time I am motivated by more than a vague interest in the field of digital preservation. I am a Board Member of the Digital Preservation Coalition in the UK; I am responsible for delivering the entire digital strategy of the National Library of Scotland and I am one of the handful of people the Scottish Government comes to when it wants advice on the future of digital preservation in my country. I also have a background in developing shared services and I have first-hand experience of how difficult it can be to turn the desire for collective action into a reality.

2. WHAT SORT OF CONSORTIUM IS APPROPRIATE FOR SCOTLAND?

The national structure of digital preservation is currently being debated in Scotland by the cultural institutions themselves, by the Scottish Government and within the digital preservation community. The philosophical battle as to whether Scotland's cultural institutions should engage in digital preservation has been won. There is clear consensus that the answer should be "yes", but the institutions are at a cross-road about the "how". Without the right vision and leadership the outcome of this consensus is likely to be a messy series of independent initiatives. This paper looks at the choices open to the cultural sector in Scotland and ponders which options for joining things up might be politically and practically feasible.

Scotland has a complex cultural heritage landscape. Digital preservation could simply be seen as an internal issue for individual organisations. However, this flies in the face of Scottish Government policies to develop shared services and to achieve economies of scale. If we collectively opt for a fragmented preservation sector, it will make achieving the goal of unified search across collections much harder. Users benefit hugely if they can delve into many collections in a single search. The principal of unified search is also aligned with the desire of the Scottish Government to make access to services both digital and easy, a policy of digital first.

Digital preservation should be driven by aligning information management practices with business needs. This means having the right tools and workflows for preserving content, including accommodating any requiremenst for continuing access. Up until now in Scotland the focus within each individual organisation has been inward-looking, concentrating on ones own data. This has led to a mixture of incomplete and technically incompatible solutions across the sector.

A digital preservation consortium for Scotland could be arranged around one or more different industry groupings or dimensions.

I am one of the leaders of the National Library of Scotland. This naturally suggests that from my perspective the consortium could be library centric. However, logically it could also be specific to the cultural sector, or it could be widened to include government data. Perhaps it should not be restricted to one country or one industry. Within each of these high level groupings there are further choices.

For practical reasons the number of options which could be efficiently explored were restricted to a few dimensions. I will now look at our main options.

3. LIBRARY DIMENSIONS

For a library centric approach, the model could envisage partnerships with university libraries, with public libraries, or with commercial libraries. In each case the scale could be city-based, or it could be spread across some or all of Scotland. Geographicaly it could also include other parts of the UK or even extend internationally. The reality is that the National Library of Scotland already participates in collaborative groupings within the library world at each of these scales for many different library purposes. However, with one very specific exception, it does not currently do so for digital preservation.

The exception is the shared Digital Library System (DLS) created to handle electronic Legal Deposit in the UK and Ireland. This common infrastructure is owned equally by the British Library, the National Libraries of Scotland and Wales, the Bodleian Library at the University of Oxford, Cambridge University Library and Trinity College Dublin.

The DLS is fairly new, only entering operation in April 2013. It includes all non-print publications and an annual copy of the .uk web, but excludes websites that are mostly made up of moving images and sound. Over 1 billion URLs have already been collected as well as more than 300,000 journal articles and many other e-resources. As is the case for print legal deposit, UK legislation dictates that electronic legal deposit can only be accessed on the premises of a legal deposit library. The DLS is managed from the British Library and features a full digital preservation environment with all content mirrored across four sites. To ensure the integrity and safety of the legal deposit holdings, it effectively operates as a walled garden. At present the work flows are integrated with the British Library's own systems. These flows are quite different to those of the National Library of Scotland, so they are not likely to offer Scotland a preservation development path for its other collections.

4. UNIVERSITY DIMENSIONS

Scottish universities and their libraries make an interesting potential pairing for several reasons. The nation has two research institutes of international stature working in the field of digital preservation. They are the Humanities Advanced Technology and Information Institute (HATII) at the University of Glasgow and the Digital Curation Centre (DCC) at the University of Edinburgh. Through these bodies, as well as through the Digital Preservation Coalition which is partly co-located at HATII, Scotland is plugged into a wide variety of national and international research preservation initiatives, all of which must benefit the National Library of Scotland and other players in Scotland.

Most Scottish universities are well advanced on the path to hosting research data repositories and at least some repositories offer an effective preservation environment. The current strategy of NLS is to leave the management of research data repositories to the universities, but there would seem to be huge long-term potential for new types of research if NLS readers could seamlessly access all of the university's research data from within the library environment.

NLS and the University of Edinburgh library currently share the same library management system sitting on the same servers at the University of Edinburgh. This was procured through the Scottish Digital Library Consortium, a member-owned co-operative. The SDLC could be a potential conduit to explore for further discussions.

However, new dimensions have recently been added to this equation. The Librarian at the University of Edinburgh, Dr John Scally, has just been appointed as our National Librarian. At the same time the University of Edinburgh has decided to migrate to a new libray management system. Helpfully, they have made the tender process consortium-friendly, so the opportunity for NLS to follow them by joining the new system is very open. We are still working through the implications of these developments.

5. NATIONAL COLLECTION DIMENSIONS

The various National Collections of Scotland provide another obvious option for building a consortium, as all of the key players are wholly or mostly financed by the Scottish Government and each body has a need for digital preservation. However, each organisation is at a very different point in the journey to create a preservation environment, giving their staff very divergent opinions of what we should do. On the plus side, three bodies are individually pushing the digital preservation agenda, though each in a different way. They are NLS, the National Records of Scotland (NRS) and the Royal Commission on the Ancient and Historical Monuments of Scotland (RCAHMS). The activities of the National Library of Scotland are detailed in other parts of this paper and will not be dealt with here.

The National Records of Scotland have quite advanced plans for building a Trusted Digital Repository. Laura Mitchell, the Deputy Keeper of the Records, recently took on the role of Chair of the Digital Preservation Coalition, offering NLS a huge opportunity for closer and more effective collaboration. We are also in talks about possible storage initiatives for physical collections.

At the Royal Commission on the Ancient and Historical Monuments of Scotland there has been sustained lobbying of government to fund the purchase of a commercial Trusted Digital Repository solution different to that of any other Scottish body. RCAHMS holds the national collections of archaeological and architectural material, as well as the second largest aerial photographic archive in the world. While awaiting a decision on the TDR, the Royal Commission has entered into a data storage agreement with the National Library of Scotland which transfers some of the preservation risks to the Library. Again there is real potential for collaborative work.

6. DMENSIONS ARISING FROM PHYSICAL INFRASTRUCTURE AND COLLABORATIVE ARRANGEMENTS

Two other factors potentially influencing the form of a future digital preservation collaborative are the existing physical infrastructure and collaborative arrangements within the sector. They manifest in three different ways.

Firstly, since 2010 the National Library of Scotland and the National Galleries of Scotland have been building a shared service for their Information Systems. As a result they now share their networks and could share their preservation environments. The politics of the shared venture have not been completely smooth and the first attempt of the National Library of Scotland to provide the Galleries with a Digital Asset Management System was rebuffed and a commercial product was procured by the Galleries. This commercial system has now been orphaned because its project funding ended, and a further merger opportunity is expected to develop soon. We are now also co-operating on a major training initiative for entry-level digitisation staff.

A second opportunity comes from the fortuitous geographical proximity of potential partners. Legislation is currently before Parliament to merge RCAHMS with Historic Scotland to create Historic Environment Scotland. Individually and together these bodies have the potential to join the NLS optical fibre network at low cost due to the close proximity of their buildings to NLS lines. The NLS optical fibre link also goes directly past the National Museum of Scotland, offering a cheap and easy connection opportunity.

Some of the other national collections are less engaged, but remain potential partners. They include the the Royal Botanic Gardens Edinburgh and the National Trust for Scotland. Their networks in Edinburgh are close enough to the National Library to make the use of commercial suppliers an efficient and economical option.

Thirdly, the National Library of Scotland has aspirations to build its own data centre, with the ultimate aim of providing the data storage and digital preservation solution to all of Scotland's national collections. This could provide a motivation for true collaboration. It is certainly an opportunity favoured by the Scottish Government. It delivers joined-up working and has the potential to be funded from sources outside the cultural portfolio as a part of building the national digital infrastructure.

7. ARCHIVAL DIMENSIONS

A similar but different grouping comes from teasing out the issues on a sector-by-sector basis. The National Library of Scotland is first and foremost a library, but it is also the second largest archive in the country, holding about 8 million archival items. In addition, NLS is currently establishing Sound Scotland, the national sound archive, and it already operates the Scottish Screen Archive as a semi-autonomous arm. This suggests that partnering with archives would be a strong option.

Sound Scotland is an interesting test case. After a nation-wide consultation in 2009, the National Library of Scotland has agreed to build a metadata repository which will provide a single central point on the web to help locate all sound archive resources in Scotland. This allows the existing sound archive structure of Scotland to remain intact, avoiding claims of any takeover of other people's assets by the National Library.

In addition, NLS is also creating a public upload facility which will ultimately demand its own digital preservation environment. The roll-out of Sound Scotland will be supported by a significant training effort to assist the people in sound archives to run their archives better and to improve their standards of preservation, both analogue and digital.

8. PRIVATE SECTOR DIMENSIONS

Some material from the National Library of Scotland and the other Scottish cultural collections has been digitised by collaborations with commercial partners such as Gale Cengage for Eighteenth Century Collections Online and ProQuest's Early English Books Online. Similarly D. C. Thomson now holds massive collections of content for the British Newspaper Archive and for genealogy material that appears in ancestary.com. In each of these cases we are relying on the preservation solutions of the commercial partner for the working copy, though the originating institution should also have the material secured in their own facility. I don't see this as an area which really offers any sort of a meaningful all-encompassing option for Scotland. It is too limited in scope. However, it does see some of our content being preserved, and that is a good thing.

9. CONCLUSIONS

Scotland is on a journey which I hope will see a viable digital preservation coalition coalesce around one or more of the dimensions discussed. Each dimension reported here has been the subject of various exploratory discussions over the past three years. Recent efforts have seen these issues being escalated from officer-level to chief executive level.

On some dimensions the funding agencies have also started to take an interest in supporting the creation of collective digital preservation solutions. This has been an evolutionary process. Over the past decade funders have done this for the storage of physical collections. As a result in Scotland it is much easier to fund collaborative building projects than it is to finance stand-alone stores. We are working to create parallel developments in the digital arena.

The National Library of Scotland's experience in building shared services with the National Galleries of Scotland, as well as its efforts to encourage the collective national procurement of a single Library Management System for NLS with the university sector, both suggest that progress will not be particularly quick, but it is a goal worth pushing for.

Then and Now: The Evolution of Digital Preservation and Collecting Requirements Over a Decade

Leigh Rosin
National Library of New Zealand
P O Box 12349
Wellington 6001
+64 4 474 3077
Leigh.Rosin@dia.govt.nz

Kirsty Smith
National Library of New Zealand
P O Box 12349
Wellington 6001
+64 4 474 3077
Kirsty.Smith@dia.govt.nz

ABSTRACT

This paper reflects on a decade of digital collecting and digital preservation development at the National Library of New Zealand. It will examine the workflows, policies and tools that have been developed in the decade since the funding for the National Digital Heritage Archive was received. The paper will look closely at the requirements that were identified for the initial development of the digital preservation system and compare them to the status of the current preservation programme and requirements roadmap.

General Terms

Preservation strategies and workflows, Case studies and best practice

Keywords

Digital preservation, Requirements, Digital policy, Ingest

1. INTRODUCTION

The National Library of New Zealand (the Library) has been actively collecting born digital heritage collections since the mid-1990s. During these first years, processes were still being developed and there were very few organisational policies governing the management and preservation of digital collections. As an organisation we were experimenting, learning and trying to figure out how to deal with these new kinds of collections.

In 2003, the governing legislation was revised, providing the Library with the legislative mandate to collect and preserve digital content under legal deposit.

The following year, government funding was secured for the National Digital Heritage Archive (NDHA) Programme. The goal for this programme was the establishment of a digital archive that would enable the Library to meet its mandate to collect, make accessible and preserve in perpetuity New Zealand's digital heritage.

The NDHA programme spent some 18 months gathering extensive business and functional requirements. In 2006, the Library formed a development partnership with Ex Libris to build a digital archive and preservation management system. The resulting Rosetta system was launched in October 2008, and for the Library the ingest and preservation of digital material became a 'business as usual' activity. To date, the archive holds approximately 5.5 million objects, spanning across 137 different formats and consisting of approximately 50TB.

Throughout the requirements and development phases of the project, the Library continued to create and acquire digital collections. Workflows, guidelines and policies developed and evolved and continue to do so, even today.

This paper will reflect on the past decade of digital collecting and digital preservation development at the Library. We will examine the workflows, policies and tools that have been developed in the decade since the funding for the NDHA was received. We will look closely at the requirements that were identified for the initial development of Rosetta and compare them to the status of the current preservation programme and requirements roadmap.

Four key functional areas will be used to drive this comparison: Ingest and acquisition of digital collections; Content maintenance; Format library; Preservation planning and execution.

Simply, we will ask the question: If we knew then what we know now, how different would our requirements and processes be?

2. INGEST OF DIGITAL COLLECTIONS

2.1 Depositing Methods

When the Library first put together its requirements for the ingest of digitally born material, they were based on a theoretical workflow model. The assumption was that content producers would 'push' digital content to us, and therefore we created an area of the system through which we could manage individual producers' details and their deposit arrangements. These arrangements would outline what they intended to provide, how they preferred to send the files and when. These arrangements would allow us to personalize the depositing experience for the content producer. Our deposit tools would be geared towards supporting this external depositing experience, and would be set up to allow producers an easy, web-based interface by which they could provide us with files and metadata.

The reality since go live has strongly tested this theoretical workflow assumption. Content producers to date have by and large preferred to make content available for legal deposit via their existing communication and/or distribution channels - websites, email subscription lists etc. They prefer to email the files or to let us know where a copy is available for download and we have had minimal uptake of the web deposit functionality. Therefore we have a "pull" rather than a "push" workflow, whereby Library staff do the bulk of the depositing for digitally born content, which continues to be acquired by the usual distribution channels.

As a result, we developed a web deposit tool as well as an area of the system where content producers could manage their submission arrangements and personalize their depositing experience, that is largely being unused by external depositors. Since library staff are doing the bulk of the depositing, we were creating and maintaining rich personalised producer accounts within our preservation system, that don't support the ingest of material and duplicate data held in our acquisitions system. Given the volume of material legal deposit staff are processing we are now moving to streamline the staff mediated workflow by associating deposits to a generic library producer, thereby avoiding the need to create and maintain individual producer records.

2.2 Ingest Tools

Since we were working under the assumption that depositing would largely be undertaken by external producers, the requirements for our ingest tools for staff did not initially include bulk, automated functionality for uploading born digital collections. Take for example our ingest tool INDIGO[1]. Our initial requirements for INDIGO were largely focused on supporting our internal digitisation programmes, who we imagined would be the primary users of the tool. While it allowed for the uploading of born digital collections as well, its main functions were tailored for allowing simple, homogenous objects (such as high resolution Tiff files), to be sent to Rosetta with minimum metadata requirements.

Several factors caused us to re-evaluate our requirements for our ingest tools over the past several years. First, as has been previously mentioned, was the "pull" nature of collecting, which was putting pressure on the Library to create easy, automated workflows for staff to use when uploading collections.

In addition, there was also a factor relating to staff confidence. In the early days of the NDHA, when the electronic deposit workflow was new, the manual nature of deposit tools was less of an issue for staff. New systems and new types of content meant there was a high degree of caution on the part of staff, and there was a desire to manually check everything that was being deposited. Therefore the need for an ingest tool like INDIGO to easily support automated workflows for depositing born digital content was minimal. However, as staff have become more experienced and increased their technical knowledge, their needs and requirements began to change. They became more confident both in their abilities and in the preservation system and they began to shift their requirements. They became interested in exploring tools and workflows that would result in the largest amount of files being uploaded with the minimum amount of manual intervention from staff.

Finally, the more staff worked with the tools they had and expanded their knowledge about digital collections, the more they were able to imagine and articulate how such ingest tools could be enhanced. Working with INDIGO for several years allowed staff to evaluate the areas of the tool that worked well, and the areas where the tool could be improved to create more automation and efficiencies. Staff were able to see more clearly how the tool could be changed to allow for more complex objects to be loaded; how functions could be altered to allow for more varied metadata inputting. They were also beginning to see how staff time could be saved by engineering the tool to do the bulk of the work. These are ideas and requirements that grew out of staff experience and have resulted in five new iterations of INDIGO being developed since 2009.

2.3 Ingest Activities and Documentation

The ingest and technical analysis activities performed during the acquisition of born digital collections, as well as how we document those activities, is another area where our policies and processes have changed greatly over time.

In the early days of acquiring digital collections, there was very little use of tools or digital forensic technologies. Our main goal was to migrate files off original media and get them onto a secure server. However, the methods used were pretty basic. For example, when files were copied there were no checksums generated before and after copying, therefore making it difficult to ascertain whether changes to the files had occurred during the transfer process. This is an area where the Library's processes are being re-developed, to ensure the authenticity and integrity of the files can be maintained throughout the transfer process.

Several tools[2] are being trialed by staff in an effort to make improvements to our processes. A fixity policy for the library, which will govern the handling of digital collections, has also had an impact. Although at the time of writing this article the policy has not yet been signed off, staff are preparing for it already. We need to prepare our processes so that they can meet the main fixity policy goal, which is "[t]o ensure that all content under the control of the Library or Archives can be, and is monitored for corruption and unauthorised change.". [7]

The documentation of ingest and technical analysis activities is another area where many changes have occurred. Staff have always described their activities, but ingest reports and file listings were sometimes missing key details about hardware/software used, actions taken or methods trialed. There was a lack of consistency and it was often difficult for staff who would later work on these collections to know what tools had been tried, what actions had been taken and why.

This continues to be a challenging area, where considerable improvement is still necessary. Although reporting is more consistent now (templates are used by all staff) and key details and actions are better documented, the process continues to be extremely labour-intensive and manual. Staff continue to try and improve the way they document their activities to make the process more efficient.

3. CONTENT MAINTENANCE

One of the Library's initial business rules underpinning system requirements was that objects would not be 'touched' prior to ingest. Any maintenance actions that needed to be undertaken would only be done within the confines of the preservation system, where they would be auditable. The initial data migration quickly highlighted the constraints of such a business rule. Our data was far less 'clean' than we imagined. During ingest, files

[1] INDIGO is an internal submission tool, developed by the Library, to integrate with the digital preservation system. It is a desktop application used by various business units to create deposits of files and metadata and upload them to the Rosetta digital preservation system

[2] Shotput Pro, FreeCommander and TeraCopy are several examples

are run through a validation stack, where a series of technical checks are run (virus checks, fixity, format identification and validation). Where we had imagined 'unclean' files would be the exception, it was immediately apparent that a large percentage of our data triggered errors in the format identification (DROID) and validation (Jhove) tools. While system requirements quickly evolved to enable rules to be set to ignore certain tool errors, not all errors were ones that we wanted to ignore.

While files with missing or wrong file extensions can be 'ignored' and ingested into the preservation system, the Library felt it was preferable to load files in a relatively clean and stable state. Once files are in the preservation system, there are limits to the actions you can perform and the tools you can use. The Library has found that for collections with large numbers of files requiring a series of fixes to be applied, it is easier and more efficient to perform these actions prior to upload. As a result the Library adopted a pre-conditioning policy; this sets the limits of change that can be introduced to digital content from the time it is brought within the control of the Library to its acceptance into the preservation system. Three key operating rules underpinning the policy are:

- Changes cannot be made on the intellectual message of the object,
- All changes must be reversible and,
- All changes must have sufficient documentation to demonstrate the reasons they were undertaken as well as a system-based provenance note that clearly describes the change that has been made to the file. [6]

Throughout its implementation of the preconditioning policy, the Library has been rethinking its requirements in respect of provenance data. Metadata in Rosetta largely conforms to the PREMIS model, and thus Rosetta's current data model supports provenance event data. However the information we require in the provenance note to fully satisfy our preconditioning policy is at a level that is more detailed than is allowed by existing metadata elements. As a result, the Library made a recommendation to the PREMIS Editorial Committee for the inclusion of a more granular Provenance metadata element. The PREMIS Editorial Committee added a new semantic unit as a place where such information could be stored in whatever structure an institution requires.

Thus, over the past decade, the Library has eased its stance on performing activities on files prior to their ingest into the preservation system and created policies and workflows to support this position. This has also resulted in updates to certain metadata elements.

4. FORMAT LIBRARY/REGISTRY

One area where the Library's requirements have changed significantly is the Format Library. When our requirements were first compiled, the Library had a fairly simple understanding of what the role and scope of a Format Library should be. The main requirement was for a library that would document formats, and link that information with supporting applications in order to identify preservation risks. Requirements were based on the assumption that most of the detailed format information would be drawn from existing registries, principally the National Archives UK's PRONOM database.

Over the last five years, we have gained a great deal of experience through interaction with collection items and use of external information resources such as PRONOM. [1]

These experiences are often mediated through tools developed by the community, thus giving us further insight into gaps or failings. All of these experiences have highlighted the "need to be able to more accurately define formats, relate them to relevant specifications, define their supported characteristics, and combine these things to form profiles that can be linked to software applications."[8] A greater understanding has led to a much more complex set of requirements.

A priority for the Library this past year has been the development of a set of requirements for a Digital Preservation Technical Registry; one that would exist as a community resource able to be used in conjunction with any digital preservation repository. This work has been undertaken under the aegis of the National and State Libraries of Australasia and has a project team comprised of the Library, the National Library of Australia, the National Records and Archives Administration (US), Archives New Zealand and the University of Portsmouth.

The Technical Registry will bring together technical information sources that currently are separate. This includes descriptions of file formats, the software applications used to create or render them, the hardware and operating systems that support the applications and files, and the perceived risks they face. It is planned that this will become the defacto hub for the rich and complex technical information and tools required to undertake digital preservation as professional activity. The Registry will benefit members of the digital preservation community through offering efficient information retrieval from one central resource, supplying trusted information and finally, supporting a community that will promote collaboration, develop best practices and peer review Registry information.

While the requirements for the Technical Registry are far more complex than those developed for a format library six years ago, a move towards a global technical registry would see a corresponding simplification at the local format library level. While a local library would have a dynamic relation to the technical registry, it is our expectation that only a relatively small subset of data would be copied to the local level. The digital preservation system retaining just enough information to support identification and reporting.

5. PRESERVATION

While preservation functionality is central to the Library's preservation system, it is the one part of the system that we have used the least. We therefore feel we are a long way from being able to specify what a fully featured preservation workflow would be like.

We are currently in the process of planning and testing preservation actions for two quite different sets of data. The first candidate, a set of Word Star files, and the second, all of the Library's web harvest Arc files. The two sets of data sit at opposite ends of the preservation spectrum: The first is a small set of files that requires boutique level preservation while the second is a large set of homogonous files requiring bulk conversion. For different reasons, both sets of data have challenged some of our original requirements/assumptions.

The Library had based requirements on the idea that the preservation system should be the primary collection tool for information required against which a preservation decision is made. However as we have worked with our curatorial staff on the Word Star content to identify the 'intellectual' aspects of the content that we need to ensure are preserved, it has become apparent that the level (both quantity and detail) of information collected as apart of a preservation plan is far greater than envisioned, and is better generated outside of a preservation system.

In the case of the large scale preservation action, converting Arc files to Warc, this has highlighted that the amount of technical data we are able to generate on the conversion process and file characteristics is far more than initially envisioned, and a lot more than can be interpreted by inbuilt technical evaluation criteria. Our original requirements for technical evaluation criteria were based on the assumption that they would be limited to only data that metadata extractors could pull out.

Another of our original assumptions that we are revisiting is the idea that, wherever possible, preservation actions should run within the preservation system. The sets of data we are currently working with have, for different reasons, caused us to rethink this idea. The Word Star preservation action is one that essentially involves handcrafting a small number of files, and it is only practical to do that external to the preservation system. In the case of the bulk conversion of Arc files to Warc, the conversion tool could be added as a simple plug-in tool and run within the Rosetta framework. However, the Library does not run a stand alone preservation system, but one that supports the day-to day collection work of the Library and the delivery of content. The Library's current system architecture and hardware infrastructure is not at the point where it can support large scale preservation processing, without impacting the performance of other system areas such as ingest and delivery.

When we first started articulating our requirements we thought we knew how we would perform preservation actions. Now that we have started to plan and test preservation actions on real content it has become apparent that we cannot as yet run transformation processes that will follow a set pattern. We do not know enough about the content in terms of all its idiosyncrasies, what acceptable change is for all types of content (or conversely what some would call the significant properties). We do not fully understand the processing requirements for each transformation and the method of undertaking it that least impacts the other library processes that depend on the system. In short, if we were preparing requirements for preservation functionality now, we expect they would be a lot simpler and less prescriptive than our original ones.

6. CONCLUSION

This has been a brief view of some of the changes in requirements that we have had since the inception of our preservation work ten years ago. Clearly, we have not included everything, nor even hinted at the scope of the changes across this decade (an entire article could be devoted for example on hardware requirements). It is clear though, that the experiences of creating initial requirements, developing a preservation system and processes and working with them as business-as-usual for five years has afforded us a different (if not better) view of what our real requirements are. We started with a theoretical, almost academic view of what we wanted our world to be. Our requirements are now shaped by business need and, as such, are focused on practical, efficient, effective processes, thus making our requirements more pragmatic.

7. REFERENCES

[1] Gattuso, J. 2012. *Throughput efficiencies and misidentification risks in DROID*. http://ndha-wiki.natlib.govt.nz/ndha/attach/ReadingResources/MSB%2BDROID%20v1_05.pdf

[2] Gattuso, J. (2012a).*National Library of New Zealand-DROID, PRONOM Developments at the National Library of New Zealand.* Paper presented at Preservation and Archiving Special Interest Group (PASIG), Dublin. http://lib.stanford.edu/files/pasig-oct2012/04-Gattuso_PASIG_presentation_2012.pdf

[3] Gattuso, J. (2012c*). Full results for DROID version 6 "Max Byte Scan" test.* http://ndha-wiki.natlib.govt.nz/ndha/attach/ReadingResources/MBSResults.pdf

[4] Gattuso, J. (2012d). *Evaluating the historical persistence of DROID asserted PUIDs.* http://ndha-wiki.natlib.govt.nz/ndha/attach/ReadingResources/Historical%20View%20of%20format%20via%20DROIDv4_2.pdf

[5] Gattuso, J. (2012e). *Main results of the DROID version tests.* http://ndha-wiki.natlib.govt.nz/ndha/attach/ReadingResources/historical%20view%20droid_results.pdf

[6] Joint Operations Group, Policy (JOGP). Department of Internal Affairs. 2012. *Digital Content Preconditioning Policy*. Pp 1-2. [This is an internal policy document, however it will be made available upon request. Contact Peter.McKinney@dia.govt.nz to make a request.]

[7] Joint Operations Group, Policy (JOGP). Department of Internal Affairs. 2014. *NDHA Draft Policies – Fixity.* P1 [This is an internal policy document, however it will be made available upon request. Contact Peter.McKinney@dia.govt.nz to make a request.]

[8] National and State Libraries of Australasia (NASLA). 2013. *Digital Preservation Technical Registry System Vision Document.* [This is available upon request. Contact Steve.Knight@dia.govt.nz to make a request.]

[9] Rosin, L. 2013. *Applying theoretical archival principles and policies to actual born digital collections.* http://ndha-wiki.natlib.govt.nz/ndha/attach/ReadingResources/Applying%20theoretical%20archival%20principles%20to%20actual%20collections.pdf

Preservation of ebooks: from digitized to born-digital

Sophie Derrot
Legal Deposit Department
Bibliothèque nationale de France
sophie.derrot@bnf.fr

Jean-Philippe Moreux
Department of Preservation and Conservation
Bibliothèque nationale de France
jean-philippe.moreux@bnf.fr

Clément Oury
Legal Deposit Department
Bibliothèque nationale de France
clement.oury@bnf.fr

Stéphane Reecht
Department of Preservation and Conservation
Bibliothèque nationale de France
stephane.reecht@bnf.fr

ABSTRACT

The scope of digital curation at the BnF covers documents digitized from BnF collections as well as born-digital material bought by the BnF or collected under its legal deposit mandate. It is therefore critical for the library to investigate if common approaches may be adopted for similar document types, whatever their origin may be. This paper proposes to focus on the case of electronic books (or ebooks), by comparing the way BnF teams intend to ensure the long term preservation of those directly digitized by the library and those that will enter through legal deposit.

Data and metadata formats are different, even though EPUB appears as the reference format for both kinds of ebooks. Acquisition procedures are necessarily specific. However, for the other steps of the treatment process, digitized and born digital books should follow similar and parallel workflows: indexing in BnF General Catalogue, access through the digital library Gallica and preservation in SPAR, BnF's digital repository. Common validation tools, characterization schemes and preservation metadata will be used in order to preserve both faces of French digital heritage.

General Terms

design, documentation, experimentation, legal aspects, performance, security, standardization, verification

Keywords

digital library, legal deposit, born-digital archives, digitization of heritage content, accessibility, ebook, DRM, EPUB, ONIX, PDF

1. INTRODUCTION

The scope of digital curation at the BnF – i.e. the set of processes intended to acquire, index, give access and preserve digital resources – covers two distinct kinds of digital material:

- on one hand, documents digitized from physical media (books, engravings, maps, etc.) held in BnF's collections;
- on the other hand, born-digital material bought by the BnF or collected under its legal deposit mandate.

Even though these resources are distinct in terms of acquisition method, legal status and heritage value, and even though they may depend on different organizational entities, we find in both cases similar types of documents: books, periodicals, images, audio and video.

It is therefore critical for the library to investigate if common approaches may be adopted for similar document types, whatever their origin may be. This paper proposes to focus on the case of ebooks, by comparing the way BnF teams intend to ensure the long term preservation of those directly digitized by the library and those that enter through legal deposit. It intends to see if they present similar characteristics and issues regarding long term preservation, if the same tools and the same workflows can be used, and where expertise can be shared. This paper does not take into account questions related to ebooks acquired for payment, as the topic is not yet mature enough. However, when the procedure for handling them is eventually designed, it will benefit from the double experience of digitization and digital legal deposit.

2. EBOOKS AT THE BIBLIOTHÈQUE NATIONALE DE FRANCE

2.1 Digitization: from image files to ebooks

The BnF has been involved in the field of digitization since 1992 through a digital library, Gallica, with nearly 2 million documents (books and periodicals, newspapers, engravings, manuscripts, objects, sound recordings, audiovisual material).

The BnF has recently decided to enhance the public dissemination of its digital contents through the production of electronic books, in addition to image and text modes. The aim of this new delivery format is to take benefit from some of the advantages of dedicated electronic book formats in comparison with the standard delivery formats usually offered by digital libraries (web, PDF):

- nomadic reading outside the digital library's website, on dedicated devices in a dedicated ebook format, EPUB;
- better dissemination of contents and better accessibility of digital contents for visually impaired people.

During the period 2011-2013, this effort has been embodied in two separate digitization programs: integration of an EPUB production process in a mass digitization program; and reprocessing of documents previously digitized, with the production of tables of contents and EPUBs. Digitization is either performed in-house or by a private contractor.

The implementation of this new format in the library has required much interaction between all the BnF teams involved in heritage digitization and some radical changes in the way the library considers its digitization activity:

- Intellectual selection of documents: for cost reasons, all the digitized books can't have an EPUB version. A choice must be made and selection criteria have to be defined: the librarian is turned into a publisher. In addition, the EPUB format is not suitable for all types of documents and some difficulties can arise in reconciling individual intellectual selection and the lack of flexibility of a mass digitization program, which needs to be fed with thousands of documents every year.

- Publishing: a "heritage EPUB" template suitable for accommodating a variety of types of documents has been defined, as well as an EPUB production charter. Again, the library must change its habits: it no longer produces facsimiles of heritage documents but a totally new editorial product.
- Quality Assurance: the BnF automatic input control system has evolved, in order to analyze this new format (metadata, technical requirements, etc.). A specific EPUB quality assurance team has been set up within the quality assurance section of the digitization service, to perform visual checking of EPUBs on reading devices and assessment of text quality.
- Archiving and long-term preservation: section 4.2 below describes the ingest process of the digitized books in the BnF long-term archiving system, SPAR.

In 2014, the Legal Deposit of ebooks is an opportunity to leverage the EPUB expertise acquired from the digitization programs.

2.2 Digital legal deposit: from web harvesting to direct deposit

In August 2006, an extension of the law on legal deposit mandated the BnF to collect, preserve and provide long-term access to all French online publications. Until recently, the BnF digital legal deposit team mainly focused on setting up a complete web archiving workflow[1]. Legal, heritage and technical reasons explained this choice. On one hand, websites were considered the most at-risk documents. Besides, tools and best practices were already available thanks to international cooperation. On the other hand, the ebook market was not very dynamic, both in terms of production and sales, and the production and distribution workflow were still maturing on publisher's side.

Finally, in the absence of a decree enforcing the law, the BnF was not able to ask for content distributed under payment and was limited to harvesting freely available resources.

When the decree on digital legal deposit was published, in December 2011 – it is now part of the "Heritage code" – the BnF started designing its ebooks deposit system.

This entry track is still under development, but some critical decisions have already been taken. First, a deposit system was preferred to web harvesting for ebooks. The BnF could have chosen to crawl ebooks hosting platforms, as it currently does for news websites [3]. However, direct deposit, via an FTP platform, was more appropriate to allow unitary treatment and cataloguing of ebooks. Moreover, in several cases, ebooks distributed online are not directly hosted on the website of the online bookstore, which is only the place where financial transaction occurs; the document itself or the link to download it is then sent to the purchaser by another mean (e.g. by email). In that case, web harvesting would not have been efficient.

Second, it was decided to work with ebook "distributors". In the French ebook market, the publisher is in charge of creating the book (both the intellectual content and the digital document); the digital bookseller is in charge of promoting and selling the ebook to end-users; the role of the distributor is to make a bridge between these two stakeholders: it receives the ebook, checks its format, verifies and enhances the quality of its metadata, and sends it to the bookseller.

Working with distributors appeared very quickly as the best solution as:

- there are few distributors compared to hundreds of publishers;
- the BnF benefits from a first set of quality controls performed by distributors, both for the ebooks and their metadata;
- distributors receive ebooks without DRMs; they are therefore able to send them to the library without DRMs.

2.3 Parallel workflows: using common tools for digitized and born-digital books

Finally, an internal workflow has been designed. Ebooks deposited by distributors along with their metadata will be received on a dedicated FTP platform. A first set of checks will be performed by the library, in order to ensure that all declared documents are available, to verify that data and metadata are consistent, and to validate the format of the ebooks and metadata. If the package that has been delivered passes the check, the ebook receives a legal deposit number and the distributor is informed that the document has been deposited. If not, the library requests a new deposit from the distributor.

The entry system is necessarily different for digitized and born digital books. However, for the other steps of the treatment process, the library intends that both kinds of documents follow similar and parallel workflows:

- Descriptive metadata will be ingested (and potentially corrected by human cataloguers) in the BnF General Catalogue, which indexes most published resources hosted in the Library.
- Access to digitized books will be given via Gallica, the BnF digital library; access to deposited ebooks will be given via Gallica intra muros. This is a specific version of Gallica, which is only accessible within the Library premises, and which gives access to content protected by intellectual property rights (as recent documents entered through legal deposit are).
- Preservation will be ensured by the BnF digital repository, SPAR, which is described more thoroughly in section 4 of this paper.

This choice has been made in order to avoid reinventing the wheel and redeveloping already existing tools. However, the reader's perspective and needs were also taken into account: readers would probably have been lost if forced to use two series of tools and applications. In short, BnF readers should not need to know BnF internal systems, and should not wonder if they are looking for a digitized or a born digital document before accessing it.

3. DIGITIZED AND BORN-DIGITAL EBOOKS: TECHNICAL CHARACTERISTICS

Management of ebooks, from entry to access, has thus become a strong issue across the whole of the BnF. Questions related to preservation have been particularly taken into account, as both digitization and digital legal deposit channels are intended to deliver documents that will be accessible over the long term. From this point of view, do both types of documents present the same characteristics? This raises two series of questions especially

[1] See [1] for questions related to digital legal deposit legislation; and [2] about the web harvesting workflow set up by BnF.

important in a preservation perspective, the one related to ebook formats, the other related to their metadata.

3.1 Formats

3.1.1 Digitization track

In 2011, the BnF chose EPUB 2 as support of its digitized books program, because it was the de facto technical standard for digital reading. The alternative “fixed layout” was not used because it was not yet specified at this time.

BnF ebooks have been designed to present as few problems as possible in terms of preservation:

- EPUB is based on standards and formats already mastered: XHTML, CSS, Dublin Core, etc.
- They are produced under a BnF charter: their content and structure are well known, and remain consistent over digitization programs.
- They do not include contents or formats that are potentially “risky” for preservation: multimedia files, programming code for interactivity, etc.

The next mass digitization program (2014-2017) will foster the accessibility to digital contents with the EPUB 3 format. This new version offers a wide range of accessibility mechanisms based on the semantic annotations of content. These mechanisms are all based on markups (HTML5 markup and EPUB 3 specific markup). Consequently, the risk for preservation is considered sufficiently low.

3.1.2 Legal deposit track

Historically the jungle of commercialized ebook formats is very dense, as every content or device producer has tended to create its own format (PDB, LRF, LIT, MOBI, etc.). Over the past couple of years though this density has tended to reduce, to the advantage of standardized formats (EPUB) or closed-source formats of the market leaders (such as Amazon’s KF8). Even if it does not solve every issue, this simplification is quite a relief from a preservation point of view, especially in a legal deposit context which theoretically extends the perimeter of the objects concerned to all produced and commercialized formats.

Above all, it is necessary to determine the limits of the scope of legal deposit. We tried to concentrate our efforts on files which can be easily defined as “books” in comparison with printed books. Therefore, ebooks on formats such as TXT or DOC are initially excluded, as they are closer to production formats than to diffusion formats. These formats are never to be found on commercial markets and distributors do not work with them. At the other end of the ebook channel, physical reading devices and their software won’t be concerned either by this deposit track.

Once these exclusions have been made, what is left? During the initial discussions, publishers and distributors made clear that the most frequently produced and sold formats were EPUB 2 and PDF; MOBI and EPUB 3 were in minority. Distributors add DRMs to these files or send them to international online booksellers (Apple, Amazon and Google); these online booksellers take then care of the migration into their own format.

It was agreed that ebooks with DRMs will be excluded, to permit manipulations of the files during the deposit process. Frequent copies are necessary for deposit, access and preservation processes, yet they are often prevented by DRMs. Setting up a legal deposit system of this kind of protected material didn’t seem feasible[2]. Our publisher partners agreed to that pragmatic position, which simplifies the whole process. Closed source formats won’t be deposited for the same reasons: preservation systems will probably be unable to deal with them, especially in the long term. In spite of the ambiguity of this position – in contradiction with the traditional objective of comprehensiveness of legal deposit –, it seemed more reasonable to proceed in this way in order to ensure long-term access and preservation – which are also part of the objectives of legal deposit.

3.2 Metadata

3.2.1 Digitized ebooks

The EPUB format embeds metadata (Dublin Core) to provide information about the digital publication. These metadata are exported from the BnF catalogue. Some of them have particular values in a library context:

- ID: ark[3] of the digital document in the BnF digital library (Gallica).
- Source: HTML link to this digital document.
- Relation: catalogue entry of the heritage document.

EPUB 3 version offers a richer description of the bibliographical metadata and enables the inclusion of accessibility compliance metadata in an ONIX[4] message.

But the EPUB file, as every digital object in the library, must also be characterized within the BnF IT systems:

- Version: EPUB 2 or EPUB 3?
- Format: standard EPUB or fixed layout EPUB?
- Quality: Bronze and Silver are heritage EPUBs produced by mass digitization programs, with two text quality levels; Gold are editorial EPUBs (enrichments, editorial works, etc.).
- Accessibility: does the EPUB embed accessible features?
- Production information: service provider, tools used, date of production, etc.

This information is relevant for various uses: diffusion, preservation, production, etc.

3.2.2 Deposited ebooks

The main idea when putting in place an ebook legal deposit was automation, including re-exploitation of metadata created by publishers and distributors for their own needs. The most used format for carrying this metadata in the book trade is ONIX for Books, a XML standard designed to support computer-to-computer communication of bibliographic information. ONIX files are generally completed by distributors from information provided by publishers and then sent to online booksellers along with ebook files.

- Advantages of ONIX files attached to ebooks by the distributors are their richness and precision: this information has to be exact because of its commercial purpose.

[2] About issues of preserving digital resources with DRMs, see [4].

[3] ARK (Archival Resource Key) is a persistent identifier system created and managed by the California Digital Library. See https://wiki.ucop.edu/display/Curation/ARK.

[4] ONline Information eXchange. See http://www.editeur.org/8/ONIX.

- On the other hand, quite a few of the data are useless from a librarian point of view (for example: prices for every country where the ebook is commercialized) and some important bibliographic metadata is often lacking (ISBN of the printed version is not always provided).

To make existing ONIX files as useful as possible for us, an ONIX model including the BnF specifications was defined by librarians and proposed to publishers and distributors. In the meantime, an ONIX-to-Intermarc[5] conversion was developed to allow an easy and automatic transformation of the trade information into bibliographic notices. This conversion also enables the use of this metadata for preservation needs: it will be reused within the METS file attached to the EPUB or PDF file into the Submission Information Packages (SIP).

4. COMMON APPROACHES FOR INGEST?

4.1 SPAR in a nutshell

SPAR, the Scalable Preservation and Archiving Repository, is the BnF preservation system [5]. It has been developed since 2005, and seeks to conform to the OAIS model. Its initial scope was to automate all entities that could be automated, and to offer a wide range of functions, in order to preserve various types of asset. Up to now, development was mainly concentrated on the Ingest, Storage, Data management and Administration modules.

The sets of documents to be ingested are grouped into tracks and sub-tracks (channels), according to their nature (digitized books, audiovisual files, web archives, administrative records…), to their legal frameworks, and to the way the BnF plans to manage their life cycle and apply preservation strategies. At the present time, SPAR ingests objects in four tracks: Digitized documents and associated files (except audiovisual), Audiovisual objects, Web legal deposit (ARC or WARC files), Third party storage (various kinds of files, from partners outside the institution); several others are in progress, including the Negotiated legal deposit track that will be presented in 4.3.

Each track needs a specific preingest module, because no producer[6] is able yet to deliver well-formed SIPs according to SPAR's requirements. These modules build SIPs depending on specific settings and send them to a generic ingest module, which transforms them into Archival Information Packages (AIPs).

Four levels of formats are distinguished in SPAR, corresponding to four levels of risk: stored (the most unsafe), identified, known and managed [6]. A "managed" format has published documentation, at least one reference tool and a characterization scheme. Besides, the BnF may define use restrictions depending on the producer.

Metadata for package and preservation information are contained in METS files, with PREMIS elements. Metadata for data management are expressed in RDF.

4.2 Ingest of digitized books

EPUB files aren't considered preservation copies of digitized books, but a medium for dissemination. Though, their cost and their value explain that the BnF intends to preserve them in the long term.

When they enter SPAR in the "Preservation digitization track", EPUB and adaptative[7] files are controlled and characterized. It was necessary to find a characterization tool and a characterization scheme for EPUB files. This difficulty was solved by using and adapting Epubcheck 3[8], in order to improve this software and make it a basic characterization tool. This solution is not yet completely satisfactory, and we are still looking for a characterization scheme in order to record the preservation metadata extracted by the tool. This is the reason why it can't be said yet that EPUB is a "managed" format in SPAR.

EPUB 2 and 3 files are accepted, in both standard and fixed layout for EPUB 3. This corresponds to the three kinds of ebook formats produced or to be produced soon in our digitization process. The quality level (Bronze, Silver or Gold, see 3.2.1) will not be checked, but the information declared by the digitization contractor will be preserved, as well as other production information.

Ebooks are at the moment considered as associated objects of books digitized in image mode (TIFF and now JPEG2000 files). They are described in the METS manifest with specific fileGrp use (*epub* or *adaptative*) and structMap type (*ebook*)[9]. So ebooks can't be ingested alone: they are delivered either with new digitized books or while reworking digitization (pictures are re-delivered with new OCR and EPUB files). The possibility to deliver an isolated ebook and then create a new completed version of an existing AIP is yet to be investigated.

4.3 Ingest of deposited ebooks

As explained in 4.1, a dedicated preingest module will be developed for the "Negotiated legal deposit track". The legal deposit is considered "negotiated" as the form of the delivery is negotiated between the distributor and the library, so that for example not all formats will be accepted.

This module will get the package ready for ingest, joining together the ebook file (EPUB or PDF), the original ONIX file, possibly a picture file of the book cover, and finally the METS manifest.

The role of the SPAR system in the legal deposit workflow is critical. On one hand, the preingest module will provide fundamental information to the General Catalogue, such as the ARK number of the ebook file, i.e. its persistent identifier. On the other hand, SPAR will deliver the deposited ebooks to the access platform and application.

Within SPAR itself, two channels will be set up, according to the ebook format and its components:

- Ebooks whose format is considered "managed" after all quality checks and characterization will be ingested in the first channel.

[5] Intermarc, in the family of MARC formats, is the BnF format for bibliographic metadata.

[6] In the OAIS model, the producer is the external or internal entity that produces the resource and transfers it to the Archive with the mandate to preserve it.

[7] DAISY format, used to create text or audio books for visually impaired people.

[8] https://github.com/IDPF/epubcheck.

[9] See BnF's METS profile for SPAR: http://www.loc.gov/standards/mets/profiles/00000039.xml.

- Files that are valid from a format specification point of view but which present some preservation threats will be ingested in the second channel. For example, EPUB files containing Flash or JavaScript elements are not considered managed as BnF does not have sufficient confidence in its capability to preserve them in the long term.

Thanks to the digitization track, Epubcheck has already been chosen to perform a new format check on EPUB files when they enter SPAR. But the choice has still to be made for PDF. It is currently investigated if Apache™ Tika 1.5[10] may be used in addition to Jhove: the first one to characterize the files; the second one to validate them against the results of Tika's characterization and against pre-defined profiles. This will also be an opportunity to improve other tracks containing PDF files (particularly administrative records), where Jhove is the only and imperfect tool for validation and characterization.

For the legal deposit track, XMP will probably be used as a characterization scheme for both formats PDF and EPUB. If this characterization format is considered relevant, it will in turn likely be used for the EPUB files produced by digitization. Thus, files and formats analyzes would be performed in a consistent manner. Every file of each format will be handled with the same tools and schemes, regardless of the channel or the track it belongs to. Only the application rules will differ.

Some critical choices are thus still to be done. The BnF intends to proceed on these questions during the current year, and to ingest the first deposited ebooks at the end of 2014 or the beginning of 2015.

4.4 From a digital strongbox to digital library stacks

In the current situation, that is for the digitization track as well as other tracks (e.g. web archiving), there is a fork in the document management workflows between access and preservation. SPAR is not a step between entry and access, but only one branch of the fork, separated from the access branch. This current solution is not really satisfactory, as SPAR still appears as a digital strongbox, not as BnF digital stacks.

The ebook legal deposit tracks will represent a chance to improve this situation. In this workflow, the SPAR system will play the role of the central application, as it will receive the documents from the entry step, send information to the catalogue and provide the books to the access application.

However, this architecture decision implies some challenges:

- First, SPAR will need to show a better ability to communicate with other library applications.
- Second, it should develop its capability to provide the expected files according to defined rules (for example if only one format for a specific book is requested).
- Third, the response time of SPAR must be guaranteed, because every slowdown or interruption will increase the delay between the entry of a document and its visualization.

In this way, the BnF will be able to ensure that it gives access only to documents that are already ingested in the repository. There won't be any difference anymore between what is preserved and what is offered to readers.

5. CONCLUSION

Ebooks from legal deposit and digitization tracks differ in various aspects: they have different legal statuses; they were not acquired for the same goals and for the same audiences; and BnF's preservation mandate for them is different. In one case, the BnF (or its contractor) is the producer of the documents; in the second case the BnF is only the depository.

Moreover, even though both kinds of ebooks are available in the same formats (EPUB and PDF), their technical characteristics may differ (use of JavaScript, of embedded content, etc.).

It is nonetheless possible to adopt common approaches and to leverage developments performed for one track to improve another track. Systems originally built for digitized books (Gallica/Gallica intra muros for access, SPAR for preservation) will be used for ebooks received through legal deposit. Common tools (Epubcheck, Tika, Jhove) and common characterization schemes (XMP) are applicable in both cases.

Benefiting from expertise of various teams with different backgrounds has actually been a strength: crossing points of views brought a global vision considering all aspects of ebooks preservation.

6. REFERENCES

[1] Illien G., Sanz P., Sepetjan S. and Stirling P. 2012. The state of e-legal deposit in France: looking back at five years of putting new legislation into practice and envisioning the future. In *IFLA journal*, 2012, vol. 38, n° 1. [http://www.ifla.org/files/hq/publications/ifla-journal/ifla-journal-38-1_2012.pdf]

[2] Le Follic A., Stirling P. and Wendland B. 2013. Putting it all together: creating a unified web harvesting workflow at the Bibliothèque nationale de France. [http://www.netpreserve.org/sites/default/files/resources/Putting%20it%20all%20together.pdf]

[3] Oury C. 2012. When press is not printed: the challenge of collecting digital newspapers at the Bibliothèque nationale de France ». In *Proceedings of the IFLA Preconference, newspaper section*, (Mikkeli, Finland, August 2012). [http://www.ifla.org/files/assets/newspapers/Mikkeli/oury_clement.pdf]

[4] APARSEN. 2013. *Report on DRM Preservation.* [http://www.alliancepermanentaccess.org/wp-content/uploads/downloads/2014/02/APARSEN-REP-D31_1-01-1_4.pdf]

[5] Derrot S., Fauduet L., Oury C., and Peyrard S. 2012. Preservation is Knowledge: A community-driven preservation approach. In *Proceedings of the 9th International Conference on Preservation of Digital Objects* (Toronto, Canada, October 2012). [https://ipres.ischool.utoronto.ca/sites/ipres.ischool.utoronto.ca/files/iPres%202012%20Conference%20Proceedings%20Final.pdf]

[6] Ledoux T. 2012. *SPAR: From Design to Operations*. Presentation at the Preservation and Archiving Special Interest Group (PASIG) (Austin, USA, January 2012). [https://lib.stanford.edu/files/pasig-jan2012/12B4%20Ledoux%202012_01_11-BnF-SPAR-FromDesignToOperations.pdf].

[10] https://tika.apache.org/1.5/index.html.

The process of building a national trusted digital repository: solving the federation problem

Sharon Webb
Digital Repository of Ireland (DRI
National University of Ireland, Maynooth
Co. Kildare, Ireland
sharon.webb@nium.ie

Aileen O'Carroll
Digital Repository of Ireland (DRI
National University of Ireland, Maynooth
Co. Kildare, Ireland
aileen.ocarroll@nuim.ie

ABSTRACT

The Digital Repository of Ireland (DRI) is building an interactive national trusted digital repository for contemporary and historical, social and cultural data held by Irish institutions. It will provide a central Internet access point and interactive multimedia tools, for use by the public, students and scholars and inform national policy for digital preservation and access. In 2011/2012 DRI conducted a requirements analysis of stakeholder needs [1].· This paper focuses on how aspects of this requirements analysis are translated into technical and policy solutions. We address how the project consortium, comprising six academic institutions, integrates with existing partner repositories and how the Digital Repository of Ireland tackles issues of repository federation in terms of storage, deposit and the legal frameworks associated with these activities.

General Terms

Infrastructure, communities, preservation strategies and workflows, case studies and best practice

Keywords

Requirements, policy, storage, deposit, user roles, use case, legal frameworks.

1. INTRODUCTION

The Hope consortium, tasked with building a federated repository of social history archives, detail a number of suggested benefits to adopting a federated model. They argue that users are less likely to turn to local catalogues to find content and that federation is more responsive to user needs. Clustering of content increases connections and links between content located in different collections, both at the national and international level, which enhances the contextual information about the digital object. Federation drives the adoption of open source solutions and shared standards which both increases the sustainability of the technical systems and the discoverability and quality of digital objects [2]. However, federation is not without its challenges. In this paper we outline how the Digital Repository of Ireland (DRI) has responded to the challenge of federation by discussing federation at the levels of storage, access management, and organisational structure.

The Digital Repository of Ireland (DRI) is building an interactive national trusted digital repository for contemporary and historical, social and cultural data held by Irish institutions; providing a central Internet access point and interactive multimedia tools, for use by the public, students and scholars; and is seeking to inform national policy for digital preservation and access. The DRI research consortium comprises six academic partners: Royal Irish Academy, National University of Ireland Maynooth, Trinity College Dublin, Dublin Institute of Technology, National University of Ireland Galway, and the National College of Art and Design. DRI is a four-year exchequer funded project (funded by the Higher Education Authority PRTLI Cycle 5), and is collaborating with Irish cultural and social institutions such as the National Library of Ireland and the Irish national broadcaster RTÉ.

In parallel with a comprehensive requirements specification phase we have developed a lean repository prototype, and published a national report [3] with the findings from our nationwide programme of stakeholder interviews to determine the digital preservation and access practices in cultural institutions, libraries, higher education institutions and funding agencies. We are working to raise awareness of the need and benefits of digital preservation and open access, while respecting and acknowledging ownership, copyright, intellectual property rights, privacy and confidentiality.

In 2013 DRI carried out a mapping exercise, examining the range of institutions tasked with caring for digital content [4]. It is possible to classify three different architectural approaches to caring for digital content:

1. Single-site repositories, in which the technical and organisational function are located in one place (excluding off-site backup). The single-site approach is often adopted by national infrastructures.
2. In 2007–2009 a number of metadata aggregators were established. This approach brings together (aggregates) the metadata of a number of single-site repositories, thus increasing user awareness of content held in various repositories.
3. Since 2009 there has been a demonstrated shift towards the establishment of multi-site repositories, in which the technical infrastructure is federated across a number of repository sites. The Internet Archive and Dataverse were early adopters of such a multi-site approach.

The first challenge faced by DRI was how did we fit into this repository landscape and in particular how would we interpret our commitment to federation.

2. FEDERATION

Federation may refer to an organization or group within which smaller divisions have some degree of internal autonomy. It can occur at a number of layers in the software and hardware infrastructure as well as at an organisational level. The OAIS reference model describes Federated Archives as “a group of Archives that has agreed to provide access to their holdings via one or more common finding aids”[5]. In this context they define a Global Community as “an extended Consumer community, in the context of Federated Archives, that accesses the holdings of several Archives via one or more common Finding Aids” [5]. Different types of federation are evident among those caring for digital content [4]. For example, Europeana is an example of a system in which DIPs (dissemination information package) containing the finding aids from each OAIS are ingested into the Common Catalog [5], it federates at the metadata layers and requires members to adhere to its standards. In contrast, the Institute for Qualitative Social Science (IQSS) model of federation is that it delegates the access controls to it’s users, the systems are primarily located at IQSS’s data centres [6], The IQSS provides the tools and infrastructure for contributors of data and meta-data and lets the ‘user’ decide on its own level of autonomy and trust. Of the three types of federation outlined in the OAIS reference model, DRI is closest to a Global Site structure, that is

> Global access is accomplished by the export of a standard-format Associated Description to a global site. The global site independently manages a set of descriptors from many Archives and has finding aids to locate which Archive owns a collection of interest. The Consumer is given a combined view of the holdings of multiple sites, which is maintained centrally. To view details of the documents, the user must access the site that contains the actual document. This is made easier when sites and clients support a standard set of protocols. [5]

In seeking to future-proof the DRI infrastructure, and in line with emerging trends, we have adopted a federated architectural approach for the DRI. In addition to the benefits outlined by the Hope project above, this also enables us to partner with existing and future digital archives, which we view as essential for a richer user experience, and to truly achieve our national mandate.

3. STORAGE

Federating at a storage level brings with it obvious advantages. DRI is building a trusted digital repository; it is a requirement of this trusted system to have high level availability (that is, with limited, controlled, downtime) and redundancy (duplicate copies of data available). Therefore, we are federating at redundancy and backup level. This approach fulfills a number of important business requirements, namely that the system is robust and reliable. Federated storage means that each federated member holds a copy of the repository, so if one goes down there are additional copies of the data and metadata available. This set up ensures that users have sustained access to content. This is a necessary feature from the user’s perspective, as a reliable service garners trust - it also helps to build a user base that has confidence in the service provided. However, this is federation in a shallow sense and is not the focus of this paper. Here we focus on at other levels of federation - the first of which is delegating responsibility to federated partners in terms of deposit and access.

4. DEPOSIT AND ACCESS MANAGEMENT

The access management of an infrastructure, repository or application server can often be centralised or distributed.[1] Access management depends on the level of federation of the system and the policies governing access. Access management can occur in a central manner where it is centrally controlled or it may be delegated back to the community. For example in the IQSS and Europeana infrastructures it would be up to the contributors to decide what can and cannot be accessed. The control is delegated (federated) to members of these organisations. This is the approach, informed by our requirements and policy interviews, that DRI is taking.

As discussed in our 2013 paper, “The process of building a national trusted digital repository: a user centric approach for requirements gathering and policy development” [1], our requirements analysis informed us that it was necessary to build ingest functionality to support single as well as bulk ingest. This activity gives stakeholders high levels of autonomy and control over the ingest (or deposit) process. Although DRI is federated at an organisational level, one approach could have been to allocate central resources to manage the ingest process on behalf of DRI partners. Instead we chose to build an automated process that distributed responsibility to the stakeholders. The driver of this model is to ensure effectiveness in the context of resource limitations. However, an additional benefit of this model is that it builds the DRI federation at an organisational level, since in order to deposit, depositors must also act as partners. This involves legal agreements, as well as training and skill sharing within and among the community of DRI partners.

Our online work-flow facilitates data ingestion to the repository remotely (via ingest tools) by authorised third parties, namely partners of the DRI project. For this requirement we have developed a process to authenticate individuals who wish to deposit data on behalf of their institution/archive/library, etc. and have identified a hierarchy of those “users” that may work on such ingestion processes. In order to create and populate collections in DRI, representatives from an institution (library, archive, museum, etc.) need to apply to DRI to become an Organisational Manager. Once signed up the Organisational Manager can assign different roles to staff (see below for legal frameworks).[2] Additional roles include Manager User and Edit User.

The Organisational Manager is a user who has full access rights to particular collections and who has signed the Organisational Manager Agreement (see below), as such they act on behalf of their particular “organization” (university, archive, research center, library). They may or may not be the depositor of content but they have permission within the system to create collections and grant Manager and Edit roles to preferred users. In most cases this will be a librarian and/or a professional archivist. An Organisational Manager can:

[1] Access management should however not be confused with authentication and identity management of users of a given system, these issues are not dealt with in this paper

[2] The Repository Administrator will grant Organisational Manager privileges following instruction by the DRI Director.

1. Create a new collection in which to deposit digital objects.
2. Assign Manager User (see definition below) roles to a registered user in DRI.

A challenge that we faced was that many large institutions, such as a university, often themselves had federated structures. Therefore, it is envisaged that there will be more than one Organisational Manager associated with these types of federated institutions. The role of the Organisational Manager is illustrated in the following use cases.

> **Use Case 1**: An Organisational Manager, the Head Librarian, wants four collections from the library (1798 Pamphlets, 20th Century Fanzines, 15th Manuscripts and Irish Soldier's Wills) ingested into the repository (DRI). The Head Librarian wants to assign the management of these collections to four members of staff who are individually knowledgeable of one area each. The Head Librarian assigns four members of staff as a Manager User, one for each collection/project.
>
> **Use Case 2**: The head of the Department of Sociology wants to use the repository (DRI) as their main repository for research data generated by their PhD students. The head of the department asks their administration staff to register to DRI and apply to become an Organisational Manager on behalf of the department. The Organisational Manager (i.e. the admin. staff) is the point of contact for all PhD students who want to deposit their research data into DRI. The Organisational Manager will create a new collection for each student and assign him or her a Manager User role.

The role of a Manager User therefore reflects the need to allocate or grant responsibility for the day-to-day management or maintenance of a collection. An Organisational Manager automatically inherits the functionality or capabilities of the Manager User and can chose to delegate or not. A Manager User is a user who has manage permission on a particular collection or collections. Although strictly speaking, this is a permission-based role, it can be thought of as a distinct user type. These user permissions should, however, be interpreted as applying only with respect to the specific collection or collections on which the user has manage permissions.

A Manager User is an authorised user who can ingest content into collections, which an Organisational Manager has assigned to them. A Manager User can manage a number of collections. They have permission to:

1. Set the metadata standard for the collection.
2. Edit the collection title.
3. Provide a description of the collection.
4. Upload funding and partner logos related to the collection.
5. Assign and remove Manager User roles.[3]
6. Assign and remove Edit User roles.
7. Set and edit access permissions.
8. Review a collection.
9. Publish a collection.
10. Review collection activity.
11. Create folders

Importantly, a Manager User must "review" a collection (e.g. access permissions, metadata, etc.,) before a collection is "published" and visible on the DRI repository. This step is both a quality review for the Manager User and a chance to ensure that access permissions are correct in cases where a Manager User is relying on an Edit User to upload content. The Manager User automatically has the same permissions as an Edit User (see definition below).

The role of the Manager User is illustrated in the following use case:

> **Use Case 3**: A librarian is assigned as a Manager User and given access to the "1798 Pamphlet" collection. They write a description of the collection to give contextual information to the project and upload their institutional logo. There are 10,000 digital objects in the collection, each of which consists of the digital asset (the image) and a metadata file (Dublin core in XML). The library has two interns to help ingest the collection into DRI, the Manager User assigns these interns the Edit User role.

Finally, an Edit User is an authorised user who can ingest content into collections they have access to. An Edit User has limited functionality/permissions but must also adhere to the deposit terms and conditions (see legal framework below).

They have permission to

1. Ingest digital objects (asset and metadata) into the repository. They can use the single ingest web form or the bulk ingest tool (currently a command line tool).
2. Edit object metadata
3. Delete unpublished objects
4. Set a collection from draft to "for review" by a manager user.

The role of the Edit User is illustrated in the following use case:

> **Use Case 4:** The library's summer intern is allocated the Edit User role by a Manager User to help ingest objects into a collection. The collection is publically accessible and contains no objects that are restricted or sensitive in nature. The Edit User uses the single ingest web form to upload objects into the repository and creates the metadata upon ingest.

DRI have developed the above user hierarchy to facilitate the various institutional constraints. It supports the distribution of work and effort when users deposit data into the DRI repository. Each user type described above can ingest into a collection for which they have access and ingest permissions. As such at any given point an Organisational Manager, a Manager User or an Edit User may be a depositor of a collection.[4] Therefore, it is important that each of these users confirm that they agree to the terms and conditions of the deposit agreement.

[3] This functionality allows the Manager User to delegate responsibilities to staff, however, we are currently reviewing whether the remove Manager User functionality should remain with the Manager User or rest solely with the Organisational Manager.

[4] A Depositor is an authorised user who can ingest objects into a collection. A Depositor may be a Organisational Manager, a Manager User or an Edit User. An Edit User cannot set access permissions to a collection or digital objects.

This user hierarchy supports the automated system that DRI have developed to ingest content from DRI partners. This automated system introduces a number of issues in terms of, "trust" - DRI partners trust DRI to hold, make available and preserve their content, while DRI must trust that depositors will adhere to the deposit agreements and in particular set the access controls on their content. Trust is introduced and based here on social and political relationships, which are then codified in a technical solution and a legal framework addressed in the next section.

5. LEGAL FRAMEWORKS

As noted above, at an organisational level, DRI is a consortium of six academic partners. Partners, in the main, not only contribute to the building of the repository at technical, policy and business levels, but also populate the repository with digital objects through demonstrator projects [7]. These demonstrator projects serve to test the repository as well as populate it with content. DRI is following the ISO 163163 (the ISO standard pertaining to Trusted Digital Repositories (TDR)) in the development of its policy framework. This standard mandates that deposit of data must take place within a specific legal framework of agreements between the repository and those who deposit -

> 3.5.1 The repository shall have and maintain appropriate contracts or deposit agreements for digital materials that it manages, preserves, and/or to which it provides access [5].
>
> The repository shall have contracts or deposit agreements which specify and transfer all necessary preservation rights, and those rights transferred shall be documented [5].

DRI faced two related challenges in developing the legal frameworks attached to deposit, access and re-use of data. Firstly, how to manage deposit licences in a federated structure and secondly, to what extent the system could be automated if paper trails or signed documentation was required.

The demonstrator projects allowed us to test the legal frameworks developed. Traditionally repositories take data from the depositor, ensure that a deposit agreement is signed and from there manage preparation and ingest of the data to the repository. There are two actors involved in this process; the depositor and the repository. Yet, as we have seen, DRI has an organisational structure that is distributed - that is, deposit in the main will not be managed by DRI personnel but instead by the depositing organisation. In many cases the depositor will not also be the owner of the data (e.g. an institution, such as a library, may be depositing data to DRI that is owned by a third party). However, the depositor will have permission from the original owner to re-use the content.

DRI is managing the distributed nature of deposit through an interconnecting network of legal agreements. Current DRI partners have, via the existing legal frameworks, the ability to assign staff to Organisational Manager roles. However, it is envisaged that DRI will expand to include new members, depositing new data. An *Organisational Manager Agreement* is an agreement between DRI and a DRI member organisation. The Organisational Agreement is attached to the Organisational Manager role and delegates responsibility for managing ingest to this user type. In contrast, the *Deposit Terms and Conditions* are attached to the collection being deposited within the archive. Either the Organisational Manager or, more likely someone they nominate, deposit the digital objects and thus have the responsibility of agreeing with the *Deposit Terms and Conditions* (discussed below).

In developing these agreements, and being mindful of the ISO 16363 standard for Trusted Digital Preservation, we encountered a number of issues that needed to be resolved. Firstly, what indeed constituted a "legitimate" deposit agreement? We noted that ISO 6363 required that "contracts and formal deposit agreements should be legitimate; that is, they need to be countersigned and current"[5] and that in most of the archives and repositories we surveyed, deposit agreements were indeed paper documents counter-signed by both parties. Instead we were proposing the use of a 'click-wrap' agreement, that is

> an agreement, formed entirely in an online environment such as the Internet, which sets forth the rights and obligations between parties. The term "click-wrap" is derived from the fact that such online agreements often require clicking with a mouse on an on-screen icon or button to signal a party's acceptance of the contract [19].

After legal consultation we were reassured that a 'click-wrap' license was as valid and legitimate as more traditional legal agreements, indeed 'legitimate' had no particular meaning in Irish contract law.

The second challenge we faced was, did we need the *Deposit Terms & Conditions* to explicitly state the access conditions, contact details and licenses attached to the deposited digital objects (as is traditionally the case) or could we transfer these responsibilities to the depositor? The Organisational Agreement outlines both organisational responsibilities and DRI responsibilities. Many of the issues covered by the Organisational Agreement are familiar to those utilised by single-site archives. From the organisational perspective there is a requirement that the digital objects deposited meet the repository documented standards (including but not exclusively those pertaining to licensing, metadata and formats), and that the repository is granted the right to make available the digital object and process them according to established data protection practices. In return, the repository undertakes to preserve the digital objects and maintain their long-term usability in accordance with the repository's preservation strategy. In addition, the agreement allows the Organisational Manager to authorise users to act as depositors, adding or modifying data within the system. The ISO 16363 requires appropriate contracts or deposit agreements. They suggest:

> An agreement should include, at a minimum, property rights, access rights, conditions for withdrawal, level of security, level of finding aids, SIP definitions, time, volume, and content of transfers [5].

DRI departs from traditional practice in that the *Organisational Agreement* states that the Organisational Manager will ensure that the appropriate access permissions are set per collection and/or object basis as applicable, that the appropriate re-use licence is set per collection and/or object basis as applicable e.g. CC-BY, etc., that any embargo dates (e.g. if the collection publication date should be delayed) are set on a collection, etc. The role of the *Deposit Terms and Condition* in this distributed system is not to record the conditions under which the repository may distribute data, rather it places responsibility on the depositor to apply these conditions themselves when depositing data. The ISO 16363 framework allows for responsibility to be placed on depositors, for example

> Agreements may place responsibilities on depositors, such as ensuring that Submission Information Packages (SIPs) conform to some pre-agreed standards, and may allow repositories to refuse SIPs that do not meet these standard [5].

In a repository which is federated at an organisational level, the depositor is delegated a much greater level of responsibility. This responsibility is captured in the various legal documents and agreements that DRI partners must agree and adhere to in order to participate in our federated system and organisation.

6. SECTIONS

The HOPE Project outlined many of the advantages of federation. In the most obvious way, federating technically at the storage layer facilitates robust and reliable back-up - this is reflected in DRI's approach to storage. This paper highlights other domains at which federation can occur. In particular DRI have developed workflows that provide a degree of internal autonomy to DRI partners - they are responsible for managing deposit of and access to their data. They have autonomous control of their data for all actions with the exception of hard delete (this is currently in discussion). Trust is embedded in contractual agreements and in the provision of appropriate training and skill development. To this end we have developed metadata user guidelines and fact sheets on formats, copyright, metadata and hosted a number of workshops and seminars. A key advantage of delegated responsibility is it drives sharing and interoperability. The delegation of control is only possible when accompanied by shared standards and protocols, however, these are not developed *by* the repository *for* depositors, rather they are created *by* the federation, *for* the federation. Our 2013 article on the process of requirement gathering and policy development concluded that “Building an infrastructure should not be considered a series of linear steps but rather a process of discussion and engagement.“ Most of the partner organisations have pre-existing repositories whose autonomy they wish to retain, yet they also need support for the task of long-term digital preservation and are cognisant of the benefits of building links between the collections they hold and collections in other partner institutions. The technical, organisational and legal infrastructure developed by DRI is responsive to the needs of our community - however it has the additional benefit of strengthening and supporting that community through the federated structures that encourage the development of shared infrastructure, policy and advocacy.

7. ACKNOWLEDGMENTS

We would like to thank Associate Professor Eoin O’Dell at Trinity College Dublin for his past and continued assistance on the various legal issues discussed here.

8. REFERENCES

[1] O’Carroll, A. and Webb, S. 2013. The process of building a national trusted digital repository: a user centric approach for requirements gathering and policy development. In *Proceedings of the 10th International Conference on Preservation of Digital Objects* (Biblioteca Nacional de Portugal, Lisboa). DOI - http://purl.pt/24107/1/

[2] HOPE (2012) Best Practices for Trusted Digital Content Repositories Best Practices for Trusted Digital Content Repositories http://www.peoplesheritage.eu/pdf/D2-4-Grant250549-HOPE-BestPracticesTrustedDigitalContentRepositories2-0.pdf Accessed 24th March 2014

[3] O’Carroll, A. and Webb, S. 2012. Digital O’Carroll, A. and Webb, S. (2012), Digital archiving in Ireland: national survey of the humanities and social sciences. National University of Ireland Maynooth. DOI: 10.3318/DRI.2012.1 available at http://dri.ie/digital-archiving-in-ireland-2012.pdf

[4] O’Carroll, A., Collins, S.,Gallagher, D.,Tang, J., & Webb, S. 2013. Caring for Digital Content, Mapping International Approaches. Maynooth: NUI Maynooth; Dublin: Trinity College Dublin; Dublin: Royal Irish Academy. DOI: 10.3318/DRI.2013.1 available at http://dri.ie/caring-for-digital-content-2013.pdf

[5] CCDS (2012) *MODEL FOR AN OPEN ARCHIVAL INFORMATION SYSTEM (OAIS) Magenta Book, The Consultative Committee on Data Systems*, p1-11 http://public.ccsds.org/publications/archive/650x0m2.pdf Accessed 20th March 2014

[6] Institute for Qualitative Social Science, http://www.iq.harvard.edu/ (accessed 25 March 2014).

[7] National College of Art and Design: *Kilkenny Design Workshops*, NUI Galway: *A Visual-Audio Demonstration of Irish Language and Cultural Heritage*, NUI Maynooth: 1916 *Letters*, NUI Maynooth: *Irish Lifetimes;* TCD: *Harry Clarke Studios Archive*

[8] Buono, F.M. & Friedman, J.A (1999) “Maximizing the Enforceability of Click-Wrap Agreements”, J*ournal of Technology, Law and Policy*, 4:3 http://jtlp.org/vol4/issue3/friedman.html Accessed 19th March 2014

A pragmatic approach to significant environment information collection to support object reuse

Fabio Corubolo
IPHS, University of Liverpool
Waterhouse Building, Brownlow St.,
Liverpool L69 3GL, UK
corubolo@gmail.com

Anna Grit Eggers
Göttingen State and University
Library, Georg-August-Universität
37070 Göttingen, Germany
eggers@sub.uni-goettingen.de

Adil Hasan
IPHS, University of Liverpool
Waterhouse Building, Brownlow St.,
Liverpool L69 3GL, UK
adilhasan2@gmail.com

Mark Hedges
Department of Digital Humanities,
King's College London
Strand, London WC2R 2LS, UK
mark.hedges@kcl.ac.uk

Simon Waddington
Department of Digital Humanities,
King's College London
Strand, London WC2R 2LS, UK
simon.waddington@kcl.ac.uk

Jens Ludwig
Göttingen State and University
Library, Georg-August-Universität
37070 Göttingen, Germany
ludwig@sub.uni-goettingen.de

ABSTRACT

When aiming to ensure the long-term usage of digital objects, it is important to carefully select what information to keep, considering also what lives outside of them. In the PERICLES project we start by analysing how such information has been described in related work, considering common definitions of metadata, context, significant properties and environment, and we come to the conclusion that we need to consider the broadest set of information, which we term environment information. Building on previous definitions, we introduce the concept of Significant Environment Information (SEI) that takes into account the dependencies of the digital object on external information for specific purposes and significance weights that express the importance of such dependencies for the specific purpose. From there we expand the definition in time considering the importance of collecting SEI during any phase of the digital object lifecycle, following the sheer curation perspective. Examples of SEI are illustrated in the very diverse use cases considered in the project, that include diverse data types from the Art domain and data from space observations in the Science domain. Finally we introduce our PERICLES Extraction Tool, that we developed to capture SEI, and present methods to extract SEI with experimental results supporting the approach. The PET tool automates the novel techniques we describe, supports sheer curation, as a continuous transparent collection process that otherwise the user (e.g. scientist, artist in our use cases) would have to find time to perform manually.

General Terms

Infrastructure, communities, preservation strategies and workflows, specialist content types, case studies and best practice.

Keywords

Digital preservation, significant properties, significant environment information, environment information, dependency graph, sheer curation, significance weight, dependency extraction.

1. INTRODUCTION

The PERICLES project (http://www.pericles-project.eu/) is an EU-funded Integrated Project focused on the problem of digital preservation. One of the areas of study is the investigation of what could constitute Environment Information (EI), in its broadest sense, in order to be able to select and capture the relevant part of that information that will sustain the use and reuse of the Digital Objects (DOs). One of the principles we have adopted is to try to explore the information based on the purpose that users have when interacting with a DO.

In Section 2, we explore relevant work and definitions of the information for the interpretation of a DO and describe our view on the subject.

In Section 3, we define Significant Environment Information (SEI) of a DO in a way that takes into account the purposes and the measure of significance of the purpose. This relates to, and in a way extends the definitions of Significant Properties (SP) of a DO. We also introduce the importance of gathering such use information in the user environment, in the sheer curation context, and describe methods to measure significance.

In Section 4 we look at examples of SEI that can be captured in the context of PERICLES case studies, in the art domain, for Software Based Art (SBA), and in the Science domain, in the scope of SOLAR experiment observations[1].

Section 5 introduces the PERICLES Extraction Tool (PET), the software tool we designed to capture SEI, and illustrate some of the techniques for environment information collection and how these can easily be adapted to different domains of use. The focus of the tool is on the context of unstructured workflows, as in many use cases users do not adopt workflow systems that can be used to analyse their flow of work.

Section 6 will describe some detailed experiments, and evaluate the results obtained using the PET tool.

Section 7 will draw conclusions and describe future work.

2. DIGITAL OBJECT INFORMATION: PREVIOUS WORK

In this section, we examine previous work on identifying and representing the information for a digital object that is relevant to support the reuse of that object, both in the long term, and across different user communities and for different purposes. We structure this examination by beginning with information that comes from the DO itself, then moving beyond the DO with the aim of identifying a broader set of information that needs to be taken into account to better support DO reuse, as illustrated in Figure 1. We recognise that this classification is one among many;

[1] http://en.wikipedia.org/wiki/SOLAR_(ISS)

the aim is however to show one thread that leads us to the topic of significant environment information, introduced in Section 3.

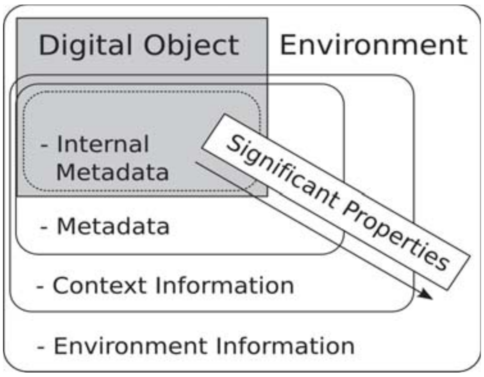


Figure 1. Our view on Digital Object and related information, from the narrowest to the broadest.

2.1 Metadata

Metadata can be defined as the information necessary to find and use a DO during its lifetime [1]. This definition covers a wide variety of information, and the Consultative Committee for Space Data Systems further refined it in their reference model for an Open Archival Information System (OAIS) [2]. This refinement covered the information necessary for the long-term storage of DOs, and they identified a number of high-level metadata categories, as follows. *Descriptive Information* (DI) consists of information necessary to understand the DO, for example its name, a description of its intended use, when and where it was created, etc. The *Preservation Description Information* (PDI) consists of all the information necessary to ensure that the DO can be preserved, including fixity (e.g. a checksum), access rights, unique identifier, context information (described in more detail in the following subsection) and provenance, which describes how the object was created. The final category arises from the fact that the OAIS manages not the DO itself, but information packages which consist of the DO as well as the DI, PDI and information required to interpret the contents of the DO (which is described by the *Representation Information (RI)*). The *Packaging Information* (PI) category describes how the information package is arranged such that individual elements can be accessed.

Standard file formats have standard structural metadata (e.g. MPEG21)[2], and *de facto* standards (e.g. the Text Encoding Initiative)[3] exist for popular formats. The situation on standardisation for the descriptive part of the RI is more complex due to the different needs of different communities, although many approaches contain the Dublin Core metadata element set [3] as a core. A catalogue of metadata standards for different communities can be found on the Digital Curation Centre website[4].

Metadata may be held internally in a DO, e.g. in the header of a structured file, or externally, e.g. in a database. Metadata may be treated as a separate entity, as it can be accessed without accessing the DO, but lack of metadata adversely affects the access to or reuse of the DO. While such information is essential for the reuse of the DO it is not in general sufficient; information concerning the external relationships of a DO, whether to other DOs, stakeholder communities, or other aspects of the environment within which a DO is created or curated, also need to be taken into account to ensure that the DO can be used fully and appropriately. This is addressed in the following sections.

2.2 Significant Properties

The concept of significant properties (SP) has been much discussed in Digital Preservation (DP) over the past decade, in particular in the context of maintaining authenticity under format migrations, given that some characteristics are bound to change as formats are migrated. The issue here was to identify which properties of an object are significant for maintaining its authenticity.

Early work in this direction may be found in [4], where SP are introduced as a "canonical form for a class of digital objects that, to some extent, captures the *essential characteristics* of that type of object in a highly determined fashion". Later work [5] investigated ways of classifying the properties: "Significant Properties, also referred to as "significant characteristics" or "essence", are essential attributes of a digital object which affect its appearance, behaviour, quality and usability. They can be grouped into categories such as content, context, appearance (e.g. layout, colour), behaviour (e.g. interaction, functionality) and structure (e.g. pagination, sections)." The concept has been adopted by standards such as [6], which describes SP as "Characteristics of a particular object subjectively determined to be important to maintain through preservation actions." Such characteristics may be specific to an individual DO, but can also be associated with categories of DOs.

An important aspect of SP is that significance is not absolute; a property is significant only relative to an intended purpose [7], or a stakeholder [8], or some other way of identifying a viewpoint. This intuition is also highly relevant to the work described in this paper.

While the concept of SP is useful for digital preservation, in its application it has usually been restricted to internal properties of a DO, for example the size and colour space of an image, or the formatting of text documents, rather than the potentially valuable information that is external to the object itself. There have been some indications of a broader conception: [5] identifies context as a category of SP, [9] refers to the need to preserve properties of the environment in which a DO is rendered, and [8] introduces the notion of characteristics of the environment. The latter associates environments with functions or purposes; this differs from what we are aiming at, which is to describe the significance of information from a DO's environment in relation to the purpose the user is following (such as editing the object, processing the object, etc.). We thus see the purpose as qualifying the significance, not the environment – a piece of information is significant for a specific purpose, but not for some other purpose.

2.3 Context

Context is a term with many definitions, a basic dictionary definition being "the circumstances that form the setting for an event, statement, or idea, and in terms of which it can be fully understood[5]". This clearly relates context to the purpose of *understanding* information, and this is a key feature of context in relation to digital objects. Context encompasses a broader range of information than metadata; it describes the setting that enables an *understanding* of a DO [10], including for example other DOs, metadata, significant properties, relationships, and policies governing the curation or use of the DO. In [11] context is defined even more broadly as 'all those things which are not an inherent

[2] MPEG21 http://mpeg.chiariglione.org/standards/mpeg-21

[3] TEI http://www.tei-c.org/index.xml

[4] http://www.dcc.ac.uk/resources/metadata-standards

[5] http://www.oxforddictionaries.com/definition/english/context

part of information phenomena, but which nevertheless bear some relation to these', where the nature of the 'relation' is left unspecified.

The OAIS model views context as the relationship between a DO (equivalent to the Content Information in OAIS terms) and its environment. In this view, the environment is considered to be necessary for using the DO, although it does not take into account two factors that we consider essential for our purposes: firstly, the possible variety of different uses to which a DO may be put, which will in general differ in the demands of 'necessity' they make on the environment; secondly, the variable strengths of the relationship with different aspects of the environment. These factors will be described and supported with examples in the following sections.

In TIMBUS [12] context is explored from the point of view of supporting business processes in the long term, describing a meta-model based on enterprise modelling frameworks. The context parameters cover a wide set of parameters, from the legal, business to the system, and technological ones, with the aim of supporting the execution of processes in the long term.

[13] presents a broad notion of context, close to our definition of environment, and recognises the importance of relationships between DOs in a collection. It further proposes a framework for contextual information that takes into account the different phases where DO context information should be gathered, and a general taxonomy of contextual entities. Automated methods for context population are also presented, but those are not overlapping the ones we are investigating in this paper, more focused on the semantic aspects of context.

2.4 Environment information

The widest set of information is the environment, which we define as consisting of all the entities (DOs, metadata, policies, rights, services, etc.) useful to correctly access, render and use the DO. The definition supports the use of unrelated DOs and conforms to the definition for environment used by PREMIS [6].

PREMIS builds on the OAIS reference model and defines a core set of metadata semantic units that are necessary for preserving DOs. The set is a restricted subset of all the potential metadata and only consists of metadata common to all types of DO. The current, published set (PREMIS2) defines a data model consisting of four entities (see Figure 2): the *Object* entity allows information about the DO's environment to be recorded amongst other information. The *Rights* entity covers the information on rights and permissions for the DO. The *Events* entity covers actions that alter the object whilst in the repository. The *Agents* covers the people, organizations or software services relevant that may have roles in the series of events that alter the DO or in the rights statements. The *Intellectual Entity* allows a collection of digital objects to be treated as a single unit.

The PREMIS working group undertook an investigation of the environment information metadata based on feedback from their user-groups that found the existing support to be difficult to use. The group reported in [14] their findings which entailed promoting the environment information to a first-class entity and not a subordinate element of the DO for the next version of PREMIS (PREMIS 3). They advocate the use of the *Object* entity to describe the environment, which allows relationships between different environment entities. This approach neatly supports the PERICLES view of the environment although PERICLES makes a distinction between the general environment and the environment significant for a particular set of purposes (termed the Significant Environment Information for a DO), which is described in the following section.

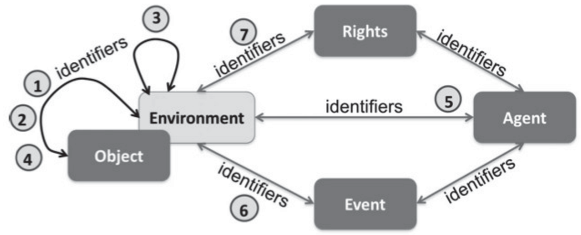


Figure 2. From [14], proposed changes for PREMIS 3 to make environment a first class object (light grey)

We consider the environment information for a DO to be the widest set of entities that a DO has the potential to be related to. This would include by definition all other DO, information, services and other information that can relate to the DO, but also other information from the environment that is not necessarily directly related to the DO but is useful for its uses (and depending on a specific use).

We consider this a wider set, although related, to the one described in the OAIS as *Representation Information*, and we take a different focus than that defined by PREMIS3. Another important distinction is that in general, we look at the environment as something defined from a DO upwards, so its definition will be related to a DO (although of course we will make sure to avoid redundancy by making good use of a linked model). In PERICLES, there is another separate view that is that of the ecosystem, that takes the view from the institution downwards. Furthermore, we consider that a part of the environment information will only be observable in the live environment of the user, so it's important to observe it in a sheer curation context as described later.

While looking at the DO environment, we consider the user an important part of it, and for that we want to observe the interaction between users and their communities, the DO, and the rest of the environment. We think that this perspective will allow us to capture the information based on the pragmatic, sometimes neglected aspects of the real requirements for making use of DO. This will also help us in the task of inferring dependencies that are not explicit and determined relevant information based on real use information.

3. SIGNIFICANT ENVIRONMENT INFORMATION

Based on the definition of environment from the previous section, we now introduce our definition of Significant Environment Information (SEI).

We define **Environment Information** for a source DO, based on the broad definition of environment in section 2.4, to be the information about the set of relationships between the source DO and any related objects from its environment.

We further define **purpose** as one specific use or activity applied to the source DO, by a given user or community. It is possible to imagine a hierarchy of purposes where a higher level purpose (as for example, 'render DO with faithful appearance' purpose) will lead to a set of detailed purposes (such as, accurate colour reproduction, accurate font reproduction etc.).

We further define **significance weight,** with respect to a purpose, as a continuous value (e.g. in the range 0-1, we have recently started refining the weights semantics) expressing the importance of each environment information relationship for that particular purpose. The significance weight will be a property of each relationship between the source DO and the DOs constituting its environment information.

Finally, we define '**Significant Environment Information**' (**SEI**) for a DO, with respect to a given purpose(s) as the set of environment information relationships qualified with significance weights. This will include both the dependency relationship (with purpose and weights) and the information that is target of the dependency.

Once SEI is determined for a collection of DOs, the different relationships can form a graph structure, where DOs in the collection could have relationship between each other (when a DO in the collection will depend on another DO in the collection for a specific purpose) or other DOs representing extracted information. This graph can be the basis to appraise (e.g. by selecting by threshold based on the weight) the set of DO that, together with their SEI, constitute the information to be preserved in order to support the selected uses of the DOs in the future.

In a less formal way, what we are aiming at is to determine "more or less all you need" when interacting with a DO for a specific purpose, and the significance of each of these information units.

Comparing this definition to that of SP to the environment, as for example described in [6], we note that the former is aimed at the collection of SEI for a DO to support the different purposes a user can have with respect to a DO, while the latter is defining the significant properties of an environment in itself. The information we aim to collect is defined by qualified relationships to other DOs, as opposed to properties of the environment.

As we mentioned in the introduction, the perspective we take is that of observing the current use of a DO before it enters a Digital Preservation system, in the systems where the DO is used, as this will allow better determination of what is significant for its use. We consider that knowing the significant information necessary to support current purposes will allow us to cover or at least know more precisely the needs also for the long term, as long as we try to support different user communities. This is because different user communities will have different purposes and different requirements, so this is a good approximation of knowing the needs of future communities (that we cannot know in advance).

3.1 Measuring significance

At this time, we are focusing on collecting a wide array of environment information, based on the relationships it can have to the DO and its estimated relevance. We are also trying to infer object dependencies that have an implied significance, by looking at use data, as described later. Still, a very relevant part of what we need is measuring the significance of the collected data. Although we don't have experimental results for now (those will come at a later stage of the project), we have clear ideas on how to define and collect it. It's in answering the question 'what for – for what purpose?' that should help us define what is significant.

Collections of data often have more than one use. Determining what information is significant depends on the use of the data. For example, the calibration of the solar measurement instrument will require calibration data, which may be a subset of the complete collection of data as well as applications necessary to read and analyse the calibration data. For a given collection not all of its environment information may be necessary for every potential use. To represent this we propose to assign weights to each relation between the collection and the environment information. The weights are based on the number of times the information is necessary for a given use. Weights will vary between 0 and 1. A weight of 1 indicates the information is essential for all intended uses of the data. Monitoring the access of information as well as regular review of the information required for each use would provide the opportunity to update the weights and could also accommodate new uses of the data.

Other factors can also be included in the weight, to express value, such as cost in time and money to collect the information as well as whether the information is proprietary (which may limit the accessibility to the information). There may also be constraints from licensing which restrict from where the data can be accessed. Any factor which influences access to the information may contribute to its weight.

Significance is also useful in the long term preservation perspective, for example to support critical analysis of the science data, as it will be a useful representation of the point of view and importance of the information for the stakeholders. It can also provide a key to understand the information.

3.2 SEI in the digital object lifecycle

In recent years, there have been various efforts within the digital curation community to establish new methods of carrying out curation activities from the outset of the digital lifecycle. A major constraint that mitigates against this is that data creators (such as researchers) typically have time only to meet their own short-term goals, and – even when willing – may have insufficient resources, whether in terms of time, expertise or infrastructure, to spend making their datasets preservable, or reusable by others (e.g. [15]). Moreover, the very volume of information that may be useful can preclude this as a practical approach, and in any case the researcher may be unaware of the utility, or even the existence, of much of this information

One approach to this challenge has been termed sheer curation (by Alistair Miles of the Science and Technology Facilities Council, UK), and describes a situation in which curation activities are integrated into the workflow of the researchers creating or capturing data. The word sheer here is used to describe the 'lightweight and virtually transparent'[6] way in which these curation activities are integrated, with minimal disruption.

Sheer curation is based on the principle that effective data management at the point of creation and initial use lays a firm foundation for subsequent data publication, sharing, reuse, curation and preservation activities, and it may be contrasted with post-hoc curation, which takes place only after the period during which the digital objects are created and primarily used.

The sheer curation model has not been extensively discussed in the scientific literature. The term has sometimes been interpreted as motivating the performance of curatorial tasks by data creators and initial users of data by promoting the use of tools and good practice that add immediate value to the data. This is, in particular, the take of [16], which discusses the role of such an approach to the distributed, community-based curation of business data.

However, this interpretation does not really address the challenges outlined above, and a more common understanding of sheer curation depends on data capture being embedded within the data creators' working practices in such a way that it is automatic and invisible to them. For example, the SCARP project[7], during which

[6] http://alimanfoo.wordpress.com/2007/06/27/zoological-case-studies-in-digital-curation-dcc-scarp-imagestore/

[7] http://www.dcc.ac.uk/projects/scarp

the term sheer curation was coined, carried out a number of case studies in which digital curators engaged with researchers in a range of disciplines, with the aim of improving data curation through a close understanding of the researchers' practice [17] [18].

In [19] the concept of sheer curation is extended further to take account of process and provenance as well as the data itself. The work examined a number of use cases in which scientists processed data through various stages using different tools in turn; however, as this processing was not carried out in any formally controlled way (e.g. by a workflow management system), it would have been impossible for a generic preservation environment to understand the significance of the various digital objects produced from the information available, as the story of the experiment was represented implicitly in a variety of opaque sources of information, such as the location of files in the directory hierarchy, metadata embedded in binary files, filenames, and log files. This was addressed by capturing information about changes on the file system as these changes occurred, when a variety of contextual information was still available, and the provenance graph was constructed from this dynamically using software that embedded the knowledge and expertise of the scientists.

The most effective way to capture SEI is through observation in the environment of creation and use of the object. We look at the interaction between the DO, the environment and the user, with time dimension. This allows us to infer dependencies that are not explicit and determine relevant information useful for use and reuse of the DO.

4. SEI IN PERICLES CASE STUDIES

The concept of SEI is now illustrated by examples in the area of digital art and space science, which constitute the main areas of interest of the use case partners of the PERICLES project.

4.1 Software Based Artworks

The following use example illustrates the SEI investigation inspired by the Software Based Art scenario from the Tate gallery. In this example a Software Based Artwork (SBA) should be migrated to a new computer system for the purpose of an exhibition. The software component of the SBA causes a strong dependency on the computer system environment. A description of SBA and an extensive study on their SP can be found at [20] and [21].

We assume there is a computer system with a validated SBA installation, which should be preserved to be able to configure and emulate the computer system environment as closely as possible for future exhibitions. The problem cannot be solved by preserving only the SBA as a DO, as the original appearance and behaviour of the software cannot be reconstructed based only on the metadata that belongs to the DO. In the context of executing the SBA's software for the exhibition are for example other dependencies such as external libraries and applications, and data dependencies (data used at run-time by the SBA). However, we have to look further at the whole environment to conceive all information that could be important for this scenario, as for example context-external running processes can affect the availability of resources, or external network dependencies. The determination, extraction and preservation of SEI are essential to solve the problem of enabling a future faithful emulation of the original system. An investigation of the environment information influence on the SP of the DO helps to identify the SEI for this use case. An example of SEI influencing the SP is when software changes the execution speed, based on the system resources, since program procedures can adapt their execution speed to the available resources depending on the programming style. This will make information about system resources SEI for the "maintaining the speed of execution" purpose. Information about display settings, as colour profile and resolution, used fonts, the graphic card and its driver is SEI that can affect the SBA appearance ('render DO with faithful appearance' purpose). Changes of programming language-related software can result in execution bugs or different speed of execution. The user interaction experience with the SBA can be affected by the peripheral driver or setting or response times that are dependent on the execution speed.

In order to determine its SEI, each SBA has to be individually analysed, regarding the use purpose and based on the properties of the artwork and the artist's beliefs regarding the SP of his artwork. Typical SEI to emulate the environment for a SBA is: information about computer system specifications, available resources, required resources, installed software and software dependencies. Other relevant dependencies to capture can be for example all the files that are used during the SBA execution, and peripheral dependencies, which can be identified by analysing the peripheral calls of the SBA. System resource requirements can be estimated on the basis of resource usage. Another example of SEI purpose, with a different set of significance weights, is when the SBA has to be recompiled because of a migration to another platform or to fix malfunctions. Here the SBA behaviour has to be validated by the comparison of behaviour patterns measured at the original system continuously in a sheer curation setting. Examples for such measurements are processing timings, log outputs, operating system calls, calls of libraries and dependent external software, peripheral calls and commands, resource usage, user interaction, video and audio recordings. The last two can be used to validate also the appearance of the artwork. If the SBA has a component of randomness, it is more difficult to evaluate its behaviour based on the measured patterns. Furthermore information about the original development environment can be useful for a recompilation, and to identify the source of a malfunction.

4.2 Space science scenario

As one of the two main use cases, the PERICLES project is considering capture and preservation of information relating to measurements of the solar spectrum being carried out by the SOLAR payload[8] of the International Space Station. The information includes operational data concerning the planning and execution of experiments, engineering documentation relating to the payload and ground systems, calibration data, as well as scientific measurements performed by solar scientists. The ultimate aim of SOLAR is to produce a fully calibrated set of solar observations, together with appropriate metadata.

We now consider three examples to illustrate the capture and use of Significant Environmental Information.

In order to validate the experimental observations of the SOLAR instrument, it is necessary to understand the impact of many complex extraneous factors on the instruments. For example, vehicles visiting the ISS can affect the trajectory of the ISS itself

[8] http://www.esa.int/Our_Activities/Human_Spaceflight/Columbus/SOLAR

and cause pollution and temperature changes. Such effects are often only uncovered by a long term analysis of the data by the scientists. Hence there is a need to capture as much of the environment as possible at the time the observations are made to enable such analysis. This includes the capture of a wide range of complex environment information relating to the instruments, the operational data, the payload sensors and events on the ISS itself. In this case, the purpose of SEI is to enable critical analysis of the solar observations by the scientists. The significance weights reflect the influence the DO captured have on the critical analysis task. These weights can change over time as additional environmental factors may be uncovered that have an impact on the scientific data. The SEI (at a given time) will therefore reflect the DOs that are relevant to critical analysis with an appropriate weighting.

In order to validate the solar measurements made by the SOLAR instrument, frequent comparisons are made with data collected independently by other scientific teams. Often the techniques and instruments are different, which provides a good way to ensure the results are not subject to unwanted effects caused by the experimental methods used. The data from other teams and the comparisons that have been made that are a valuable part of the environment metadata for the SOLAR data. The capture of the validation experiments themselves can be captured by the PERICLES PET tool and appropriate metadata created. This would include validation scripts and dependencies between subsets of the data, and would constitute (part of) the environment information. The purpose associated to the SEI is the validation of the scientific data by the science community. The significance weights reflect the value of specific data objects in the validation of the SOLAR dataset. The SEI can assist scientists in assessing the quality and reliability of the data produced.

A third example relating to the PERICLES science case study relates to the operational data for the SOLAR experiment, which is primarily created and managed by the mission operators, who operate the experiments on ISS remotely from the ground station. The operations data includes the planning, telemetry and operations logs. Given the huge complexity and volume of the space mission information, a major issue for the operators is information overload. An important task for the operators is to resolve anomalies. Anomalies occur when the normal operational parameters of the instrument are exceeded, such as overheating. Identifying and resolving anomalies often requires extensive research in the archived operations data and documentation. In this case, the digital object to be preserved is the catalogue of known anomalies and the environment information is the aggregation of all operations data. The purpose for the SEI is the identification of a specific anomaly. In this case, the significance weights indicate the relevance of a specific DO, such as a piece of documentation for the instrument or an excerpt from the archived telemetry to the particular anomaly. Thus the SEI provides a way to indicate all the environment information relevant to identifying and debugging a specific anomaly.

5. SEI EXTRACTION AND THE PERICLES EXTRACTION TOOL

Based on these premises, we are building a tool to help capture and record the environment information from the systems where the DOs are used. While different projects looked at sheer curation for very specific domains and use cases [16],[18],[19], we have built a generic, modular framework that can be adapted to support different use cases and domains with specific modules and configuration profiles. Our tool is focused on information extraction, while others target different aspects of information curation. We have also addressed the context of unstructured workflows, where the user is not adopting any workflow system, making it important to observe the flow of events in an agnostic framework.

5.1 General scenario for SEI capture

We briefly describe here a general scenario for the information capture that we are aiming at with our PERICLES Extraction Tool (PET). This should make the tool description more clear. In this scenario we observe and collect environment information from a user's computer as he interacts with DOs for different purposes. The tool is installed with the agreement and under the full control of the user. We want to look individually at the environment changes as the user e.g. calibrates some data, runs unstructured analysis workflows, creates new DOs and in different ways makes use of the data by access, interaction, and transformation.

We have different objectives that we want to accomplish, where each depends on the previous one:

1. Use the PET tool to collect environment information when the DOs are used, based on specific profiles;
2. Analyse the information to infer new relationships;
3. Assign values to the dependencies based on the purpose and significance (significance weights).

The current development status that covers mostly the first objective and starts to address the second.

5.2 The PERICLES Extraction Tool

The PET is a framework for the extraction of SEI, soon to be open sourced[9]. It can be used in a sheer curation scenario, where it runs in the system background and reacts to events related to the creation and alteration of DOs and the information accessed by processes, to extract environment information with regard to these events. All changes and successive extractions are stored locally on the curated machine for further analysis. It supports also an environment information snapshot mode, which is intended for the extraction of information that doesn’t change frequently.

The tool aims to be generic, as it is not created with a single user community or use case in mind, but can be specialized with domain specific modules and configuration. PET provides several methods for the extraction of SEI, implemented as extraction modules as displayed in Figure 3. Once PET has been configured for a particular scenario, it runs in a way that doesn't interfere with system activities and follows the sheer curation principles. An automated selection of SEI based on the use of the DOs and following the ideas outlined in paragraph 3.1 is going to be developed in a future phase.

Two different types of environment information can be distinguished: one is information directly related to a specific file, such as the location of the file at the system, or information related to the modification of the file. The other type is independent of any DO and specific to the environment, as for example general system specifications. Monitoring daemons, also displayed in Figure 3, observe the environment continuously, and trigger customised extractions based on events, as for example modifications of observed files, the creation of new files in observed directories, processing events as file openings by applications, and specific system calls. So extraction in PET sheer

[9] Apache licensed, final approval pending; source code will be available at: https://github.com/pericles-project/pet

curation mode is always related to environment changes, to avoid redundancies.

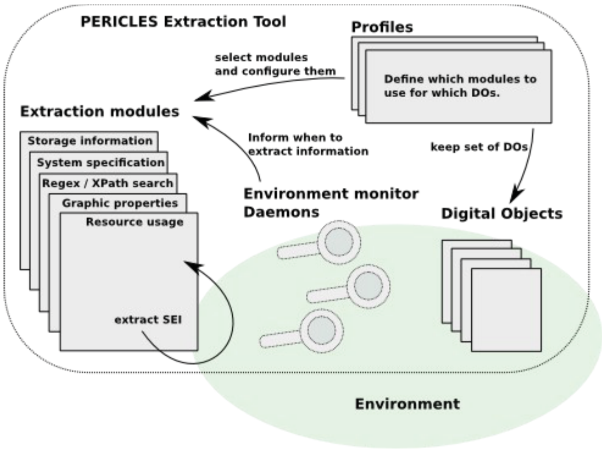


Figure 3. SEI extraction with the PERICLES Extraction Tool.

As a principle we have used existing libraries and tools, where possible, to reduce the module development times. Currently implemented information extraction modules include, among others, modules to extract:

- Available and used system resources;
- Information in files with the help of configurable regular expressions or XPath expressions;
- File format identification and checksums;
- Currently running processes;
- Event information (file and network) from processes;
- Graphic configuration information;
- MS Office and PDF font dependencies.

Furthermore we implemented generic configurable modules to execute native system commands configurable for specific needs.

With PET it is possible to create extraction profiles for different purposes to manage the information diversity. The profiles contain a set of investigated files, belonging to DOs, and a set of configured extraction modules to fit for the purpose. Future developments will include significance evaluation, as described in section 3.1, for the creation and selection of extraction profiles. Daemon modules for process and file monitoring allow the inference of process and file dependencies as described in the next section.

It is important to note that the major aim of the tool has been to enable the collection of the relevant information from the live environment, and in response to relevant events. The raw data collected will be further analysed in the tool in later tasks in the PERICLES project to conclude higher level SEI. We are also investigating techniques to encapsulate the extracted SEI together with the related DOs, to avoid data loss. These techniques will be implemented in a further PERICLES tool which will interact directly with the PET.

5.3 Extracting SEI by monitoring software

A promising technique to extract relevant information from a DO environment that we have started to develop is to look at the software currently executing on the observed system, and based on that to perform an analysis of the system calls and files used by an application. Based on a configurable set of parameters, it allows a more accurate examination of the system, and to infer dependencies between observed files based on the system activity. This will allow a reasonable amount of general information to be gathered all the time, while going in depth with the analysis of the activities when an interesting set of parameters will indicate the likelihood of a particular activity being executed.

We first describe here a simple scenario that will allow us to illustrate how such SEI collection should happen:

A scientist is calibrating data, using some specific scripts, as described in section 4.2. The PET tool is running on the scientist's machine, monitoring the environment for events that can have importance for the information collection.

The execution of a specific script triggers the event: data calibration, indicating that the user is calibrating this set of data using this script with these parameters;

Based on the event information and the state of the system the tool will first start examining the system in more detail, for example by starting a more detailed examination of the parameters and input data for the script, or observing other target applications such as office software;

A series of events and environment information is collected; this will be used to infer the activity being executed (user's purpose), and the dependencies between DOs (by looking at patterns of use, and co-occurring use of different DOs from specific software processes, dependencies based on the script, its input and output parameters; or based on other heuristics).

- By using a larger series of this collected data, we may be able to assign a significance value to the dependencies (for example by looking at how often DOs of type X is used in conjunction with DOs of type Y).
- The collected data could also be stored and kept for improving the analysis, for example by using more complex and time-consuming algorithms.

These dependencies can be mapped, automatically when possible, into a graph structure, where the edges are weighted to illustrate the importance of each dependency for the execution of an activity. The most important dependencies can then be identified defining the environment information to be extracted, on the base of the dependency graph, which helps to determine the SEI to be extracted for similar scenarios.

5.4 Provenance and other related work

Provenance information is a type of metadata that is used to represent the history of the transformations and events for a DO.

As part of our scenario, some of the data collected will be in the form of provenance information. We are exploring how such processing history of the DO can help us to infer dependency relationships, as described in more details in paragraph 5.3.

Our tool's final aim is to collect relationships between the original DO and the significant information for a specific purpose, in contrast to provenance that addresses how the DO had been created. Such dependencies are not related to single events, and are not reports of what has happened, as in the case of provenance information. It will be still possible to use our tool to collect useful provenance information, although it is not our main focus. In the development of PET, we have considered different provenance collection tools, to see if they could be helpful for our use cases.

One such example, SPADE[10], is a cross platform tool for the collection of provenance information. Its architecture [22] is similar in some ways to the one we independently designed for PET, with reporters that have a role similar to that of our modules. Spade and its modules are focused on collecting provenance information, and do not cover the variety of information we are addressing with the PET tool. We are also trying to limit the amount of information to the portion that is useful to determine SEI, and we discovered there is not a good match with the existing modules (although some of the techniques used have similarities).

The TIMBUS project [12] investigates the preservation of business processes. Although the environment information for this purpose marginally overlaps with one we are considering in this paper, TIMBUS aims at the context of the business processes, whereas we focus on assessing which environment information is, or could become, important for different uses and reuses, taking into account various purposes and informal user context while looking at the interaction with the DOs. Both our PET and TIMBUS context population tools[11] could benefit from each other by supporting the other's information extraction techniques. For example, [23] presents a different scenario: a scientist is running experiments using a formal workflow system (Taverna[12]), and the aim is that of preserving the process used by the experiment. While this is of course a worthy approach, it differs from our intent as it is based on a scenario where the user defines formally the workflow and other relevant information. In our case, we attempt to capture the process in an existing environment, where a formal workflow may not be defined.

The National Software Reference Library (NSRL) Diskprints[13] "is an attempt to comprehensively describe the changes in a computer system as a result of the influence of a software package", and could be used as an extension module to investigate in detail the state of the environment's software in its lifecycle.

6. EXPERIMENTAL RESULTS AND EVALUATION

We here describe the experiments we have set up to validate the functionality and important aspects of the framework. In all these experiments, the common steps are:

1. The PET tool is installed, configured and started on the machine where the DOs are used
2. The user interacts with the system while PET collects EI in the background
3. The environment information, DO events and changes are stored for future use and analysis

6.1 Space science

6.1.1 Operations: anomaly related information

As described in the third example of paragraph 4.2, operators dealing with anomalies usually find their solution searching through a multitude of documents. This can include for example solutions from previous anomalies, telemetry, console logs, meeting notes, emails, etc. Such data, although present in the storage, requires experience and its selection is a task that requires specific knowledge that is usually passed from operator to operator. For this reason we are addressing the collection of such dependencies between anomalies and mission documentation, in order to preserve useful information that is otherwise not captured.

In more detail, when an anomaly occurs, the issue is recorded on the 'handover sheet'. Different procedures are executed to solve the issue, and the operator's need to access the relevant documentation. We have set up a simplified experiment to show what significant environment information can be collected in this scenario. In order to support this scenario, we set up a specific PET profile that tracks the use of relevant software on specific files, using the PET software monitor; this enables us to have a trace of the documents that have been used at a given moment in time, as illustrated in Figure 4.

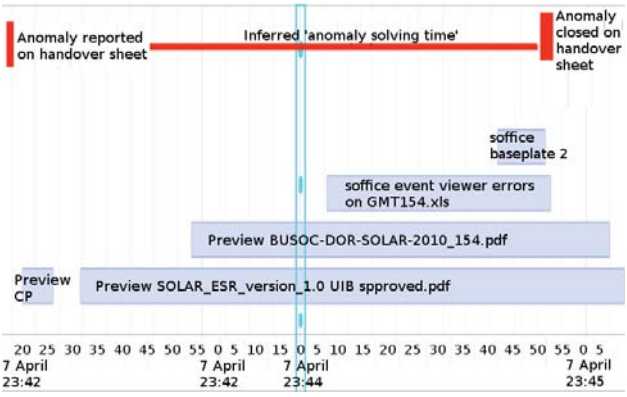
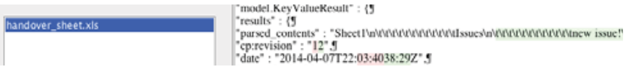


Figure 4. Trace of document use (based on open and close times) collected from process monitoring (blue) with anomaly solving time (red) collected using file change monitoring.

At the same time, it is possible to observe the 'handover sheet' and track the reporting of an anomaly start and end times (as shown in Figure 5 where a new issue is written in the document).


handover_sheet.xls

```
"model.KeyValueResult" : {¶
"results" : {¶
"parsed_contents" : "Sheet1\n\t\t\t\t\t\t\t\t\tIssues\n\t\t\t\t\t\t\t\t\t\tnew issue!"
"cp:revision" : "12".¶
"date" : "2014-04-07T22:03:4038:29Z".¶
```

Figure 5. Screenshot showing changes in the 'handover sheet' tracked by the PET tool, used to determine anomaly time

The connection between the documentation track and the 'handover sheet' tracking can allow us to infer the 'anomaly solving time span' (indicated with a red line in Figure 4) and assume there is a dependency between the solution to the anomaly and the documentation that was used between the start and end of the anomaly.

In future work we will consider more complex issues that we have ignored in this simplified example, such as the 'noise' that can be reported by the event tracking. This 'noise' can be for example due to the fact that users often multitask, so there can be unrelated documentation that was used but not relevant to the anomaly solution, or documentation that was quickly opened and closed may also indicate in some cases that the document was not relevant. We will explore also ways to obtain a fine-grained tracking, as for example to include what pages have been consulted in a document. We are planning to dedicate effort to a more careful analysis of the collected data in the next phases.

[10] https://code.google.com/p/data-provenance/

[11] https://opensourceprojects.eu/p/timbus/context-population/wiki/Home/

[12] http://www.taverna.org.uk/

[13] http://www.nsrl.nist.gov/dskprt/diskprints.html

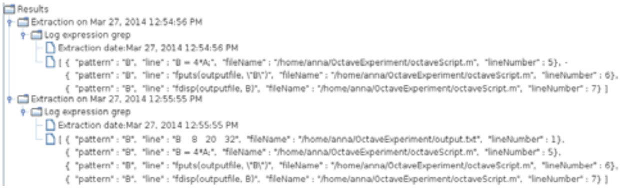
6.1.2 *Extracting results of scientific calculations*

The following experiment illustrates how SEI extraction can be useful for examining scientific calculations. This experiment uses an extraction module that extracts whole lines from files. It is configured to monitor an output directory of the open source tool GNU Octave[14] and to extract calculation results with the aid of a regular expression. The extraction module is originally intended for the extraction of particular log messages from a log directory.

The scientist uses PET to track the resulting development of an Octave-script execution over time and in relation to the script lines that are relevant for the result. This enables the possibility to understand the resulting changes in relation to script formula changes. First the user configures the module by specifying the output directory and the regular expression to search the result line, which is, similar to the name of the result variable, just "*B*" at this example. Then the sheer curation mode of PET is started, to monitor the directory, which triggers an initial extraction. At the time of this first extraction the script wasn't executed. We used the following script for this example:

```
#script for octave example
1;
outputfile = fopen('output.txt', 'w');
A = [ 2, 5, 8];
B = 4*A;
fputs(outputfile, "B")
fdisp(outputfile, B)
fclose(outputfile)
```

Then the scientist starts his normal octave workflow and executes the script. The PET detects the file changes in the configured output directory and triggers a new extraction of the selected module. The following screenshot shown in Figure 6 displays the results of the first and second extraction:



```
Results
  Extraction on Mar 27, 2014 12:54:56 PM
    Log expression grep
      Extraction date:Mar 27, 2014 12:54:56 PM
      [ {  "pattern" : "B",  "line" : "B = 4*A;",  "fileName" : "/home/anna/OctaveExperiment/octaveScript.m",  "lineNumber" : 5}, -
        {  "pattern" : "B",  "line" : "fputs(outputfile, \"B\")",  "fileName" : "/home/anna/OctaveExperiment/octaveScript.m",  "lineNumber" : 6},
        {  "pattern" : "B",  "line" : "fdisp(outputfile, B)",  "fileName" : "/home/anna/OctaveExperiment/octaveScript.m",  "lineNumber" : 7} ]
  Extraction on Mar 27, 2014 12:55:55 PM
    Log expression grep
      Extraction date:Mar 27, 2014 12:55:55 PM
      [ {  "pattern" : "B",  "line" : "B    8   20   32",  "fileName" : "/home/anna/OctaveExperiment/output.txt",  "lineNumber" : 1},
        {  "pattern" : "B",  "line" : "B = 4*A;",  "fileName" : "/home/anna/OctaveExperiment/octaveScript.m",  "lineNumber" : 5},
        {  "pattern" : "B",  "line" : "fputs(outputfile, \"B\")",  "fileName" : "/home/anna/OctaveExperiment/octaveScript.m",  "lineNumber" : 6},
        {  "pattern" : "B",  "line" : "fdisp(outputfile, B)",  "fileName" : "/home/anna/OctaveExperiment/octaveScript.m",  "lineNumber" : 7} ]
```

Figure 6. Screenshot of the PET showing a calculation result extraction

The result of the first extraction shows the locations of the scripts result variable *B* in the not yet executed script, which also lies in the observed output directory. At the result of the second extraction the line of the output file with the result variable *B* and its line number can be seen. This is the extracted result of the scientific calculation.

Since also the location of the result variable at the original script was extracted, an easier understanding of the dependencies between results and locations at the script is enabled. A continuous extraction over long periods of time makes an observation of result changes in relation to changes of the script formulas possible. The PET indeed needs highly customised configurations for the example, but these enable it to adapt to specialised scenarios.

6.2 SBA: system information and dependencies

This experiment is about collecting dependencies from a SBA, as described in paragraph 4.1. The PET tool is executed on the SBA and will extract a series of information useful for the understanding and future use of the SBA.

6.2.1 *System information snapshot*

Various information pertinent to the scenario of emulating an environment of a SBA, as described in section 4.1 can be extracted by the PET with a snapshot extraction. To these belongs mainly information that doesn't change continuously, as the systems hardware specification or installed graphic drivers.

Table 1 shows a portion of the result of a snapshot extraction executed by the PET. To the significant information belong system hardware specifications, the CPU, system graphic settings as the installed fonts and display information, and information about the operating system and development toolkits used to program the SBA's software are listed here.

Table 1. SEI snapshot extraction with the PET

Extraction Module: CPU specification snapshot	
model	Intel Core(TM)i5-3470CPU@ 3.20GHz
totalCores	4
Extraction Module: Graphic System properties snapshot	
font_family_names	Bitstream Charter", "Cantarell",...
displayInformation	isDefaultDisplay=true, refreshRate=60 ..
Extraction Module: Operating System properties	
user_language	En
os_name	Linux
Extraction Module: Java installation information	
java_home	/opt/java/jre
java_vendor	Oracle Corporation
java_version	1.7.0_15

In order to capture the type of information that changes constantly (in the SBA scenario this is mainly the use of system resources) it's possible to use PET's continuous extraction mode. A measurement of resource usage values over a long period of time can be analysed to identify behaviour patterns, which can be used to validate the correct behaviour of a new software installation. Other examples of measurements are those of CPU usages, executed by PET's *"CPU usage monitoring"*-module, whereby the changes over time can be traced.

Another example of such runtime information that can be collected and be useful for assessing the dependencies of a SBA is the file-system and network usage information (all the files and network connections used during the execution of the SBA) that can be collected by the PET tool with a specific extraction profile.

The extraction results enable the configuration and emulation of a new environment for a SBA, as described in section 4.1.

6.2.2 *Extracting font dependencies*

The PDF format gives the ability to embed the font types used in a document, to guarantee faithful reproduction of the document even when the DO is moved to an environment does not include them. It still is the case that PDF documents are created without the inclusion of at some necessary fonts (for user choice or application blacklisting). To recognize such missing external font dependencies, that are particularly relevant in the case of a PDF file used in a SBA, we created a module that will analyse PDF files and extract a list of used but not embedded fonts. This list determines dependencies between the DO and the listed fonts, relevant for accurate rendering.

[14] GNU Octave https://www.gnu.org/software/octave/

7. CONCLUSIONS, FUTURE WORK

In this paper we presented our work on determining what information is significant to collect, from the widest set of the Environment Information. We presented a definition of Significant Environment Information that takes into account the purpose of use of a DO, and can apply to relationships with significance weights. We also presented ways to determine significance weights and their relations to the DO lifecycle.

Finally, we presented the tool we are developing to collect such information, together with its methods of extraction, and showed experimental results to support the importance of such information. We believe the importance of the contribution also lies in the way that the information is collected, that is domain agnostic and aims at collection in the context of spontaneous workflows, with minimal input from the user and very limited assumption on the system structure and its infrastructure.

We plan to continue our work on exploring new methods of automated information collection, and improving the filtering and inference of dependencies. We also plan to explore and implement the methods for determining significance described in section 3.1, and look at the aspects of dependency graphs based on the purpose and significance weights that the tool will allow to infer.

8. ACKNOWLEDGMENTS

This work was supported by the European Commission Seventh Framework Programme under Grant Agreement Number FP7- 601138 PERICLES.

9. REFERENCES

[1] Press, N. I. S. O. 2004. National Information Standards Organization. Understanding Metadata.

[2] CCSDS, J. 2012. *Reference Model for an Open Archival Information System (OAIS)*. CCSDS 650.0-M-2, Magenta Book.

[3] Dublin Core Metadata Initiative. 2008. Dublin core metadata element set, version 1.1.

[4] Lynch, C. 1999. Canonicalization: A Fundamental Tool to Facilitate Preservation and Management of Digital Information. *D-Lib Magazine*, 5, 9 (Sept. 1999)

[5] Hedstrom, M., and Lee, C. A. 2002. Significant properties of digital objects: definitions, applications, implications. In *Proceedings of the DLM-Forum* 200, (May 2002), 218-27.

[6] PREMIS Editorial Committee. 2008. PREMIS data dictionary for preservation metadata, version 2.0.

[7] Knight, G. 2008. Deciding factors: Issues that influence decision-making on significant properties. InSPECT project report. *Arts and Humanities Data Service/The National Archives. At http://www.significantproperties.org.uk/ deciding-factors.html*

[8] Dappert, A., and Farquhar, A. 2009. Significance is in the eye of the stakeholder. In *Research and Advanced Technology for Digital Libraries*. 297-308. Springer Berlin Heidelberg.

[9] Knight, G. 2010. Significant Properties Data Dictionary. InSPECT project report. *Arts and Humanities Data Service/The National Archives. At http://www.significantproperties.org.uk/sigprop-dictionary.pdf*

[10] Chowdhury, G. 2010. From digital libraries to digital preservation research: the importance of users and context. *Journal of documentation*, 66,2, 207-223.

[11] Kari, J., and Savolainen, R. 2007. Relationships between information seeking and context: A qualitative study of Internet searching and the goals of personal development. *Library & Information Science Research*, 29, 1, 47-69.

[12] The TIMBUS EU project, http://timbusproject.net/

[13] Lee, Christopher A. 2011. A Framework for Contextual Information in Digital Collections. *Journal of Documentation* 67,1, 95-143.

[14] Dappert, A., Peyrard, S., Chou, C. C., and Delve, J. 2013. Describing and Preserving Digital Object Environments. *New Review of Information Networking*, 18, 2, 106-173.

[15] Perspectives, K. 2010. Data dimensions: disciplinary differences in research data sharing, reuse and long term viability: A comparative review based on sixteen case studies. *Digital Curation Centre, UK, available at: http://www. dcc. ac.uk/sites/default/files/documents/ publications/SCARP-Synthesis.pdf.*

[16] Curry, E., Freitas, A., and O'Riáin, S. 2010. The role of community-driven data curation for enterprises. In *Linking enterprise data.* 25-47. Springer US.

[17] Lyon, L., Rusbridge, and C., Neilson C., Whyte, A. 2009 Disciplinary Approaches to Sharing, Curation, Reuse and Preservation, *Digital Curation Centre, UK, available at: http://www.dcc.ac.uk/sites/default/files/documents/scarp/SCARP-FinalReport-Final-SENT.pdf*

[18] Whyte, A., Job, D., Giles, S., and Lawrie, S. 2008. Meeting curation challenges in a neuroimaging group. *International Journal of Digital Curation*, 3, 1, 171-181.

[19] Hedges, M., and Blanke, T. 2013. Digital Libraries for Experimental Data: Capturing Process through Sheer Curation. In *Research and Advanced Technology for Digital Libraries.* 108-119. Springer Berlin Heidelberg.

[20] Falcão, P. 2010. Developing a Risk Assessment Tool for the conservation of software-based artworks. MA thesis, BFH, Hochschule der Künste Bern (HKB).

[21] Laurenson, P. 2014. *Old media, new media? Significant difference and the conservation of software-based art.* In Graham, B. *New Collecting: Exhibiting and Audiences after New Media Art.* Chapter 3. University of Sunderland, UK.

[22] Gehani, A., and Tariq, D. 2012. SPADE: Support for provenance auditing in distributed environments. In *Proceedings of the 13th International Middleware Conference,* (2012 Dec.), 101-120. Springer-Verlag New York, Inc..

[23] Strodl, S., Mayer, R., Rauber, A., & Draws, A. 2013. Digital Preservation of a Process and its Application to e-Science Experiments. In *Proceedings of the 10th International Conference on Preservation of Digital Objects (IPRES 2013)* 1. Springer.

The SCAPE Policy Framework, maturity levels and the need for realistic preservation policies

Barbara Sierman
KB, National Library of the
Netherlands
PO Box 90407
2509 LK The Hague
+31 70 314 01 09
Barbara.Sierman@kb.nl

ABSTRACT

A digital preservation policy is an essential document in which an organization summarizes its approaches to achieve the goals and objectives for the long term preservation of the collections in its digital archive. In this paper the reference to preservation policies in various standards is compared with a set of publicly available preservation policies, showing that there is a big gap between theory and practice. Recent work done in the European project SCAPE (http://www.scape-project.eu/) in building a Catalogue of Policy Elements could contribute to bridging this gap. The paper concludes with suggestions to further develop the practical use of preservation policies by aligning them to the maturity level of the organization.

General Terms

strategic environment, preservation strategies and workflows, theory of digital preservation, case studies and best practice

Keywords

Preservation Policies, OAIS, TDR, TRAC, SCAPE Policy Framework

1. INTRODUCTION

A Digital Preservation Policy is an essential document in which an organization summarizes its approaches to achieve the goals and objectives for the long term preservation of the collections in the digital archive. Phrases like "*Without a policy framework a digital library is little more than a container for content"* [8], p. 68] and "*A policy forms the pillar of a programme for digital preservation* " [17], p.3] are underpinning this notion and show that the importance of preservation policies is a generally accepted one in the digital preservation community. A growing number of organizations in various disciplines see themselves faced with a mandate to preserve digital collections for the long term. This task of keeping large digital collections accessible over time is no longer restricted to libraries and archives.

Preservation policies, together with the explicitly formulated strategy of an organization, play various roles. One of them is informing the stakeholders of the digital archives about the activities. Stakeholders include the staff, the depositors and the users of the digital archives as well as the general public and the designated communities for which these organizations preserve their collections for the long term. Every stakeholder has a (different) interest in transparency and openness about the approaches an archive is choosing. This is very much related to the "trustworthiness" of the digital archive, for which such transparency is a key element. In practice, digital archives will base their daily activities on organizational policies and procedures. Making these preservation policies publicly available will better inform the stakeholders. Depositors will be able to compare digital repositories, the users will know what they can expect and staff will know how to organize their work. With a growing group of long term digital archives, one would expect that there is an abundance of published preservation policies out there. This however is not the case. For various reasons, this is a lost opportunity. Often these organizations, like libraries, archives and data centers are publicly funded and there is a growing awareness that therefore not only the directly involved stakeholders should be informed about the achievements of the organizations. Because digital preservation implies a long term financial commitment, there is a pressure on these organizations to show the value and benefits of their activities and how tax payers' money is spent. This stresses the importance of the digital preservation community to be transparent and realistic in stating the preservation policies.

But what is a good preservation policy, and what should be described in it? Are there rules and guidelines? In an attempt to answer these questions the requirements for preservation policies, as defined in the two most important standards for the digital preservation community are analyzed. One standard is the Open Archival Information System (OAIS) [34]. The other is the ISO 16363 standard for Audit and Certification of Trustworthy Digital Repositories (TDR)[2]. These two standards were input for work on policies recently done in the European SCAPE project. The Catalogue of Policy Elements that was created in this project, will be explained as well as the results of an analysis done on a set of publicly available preservation policies. Finally some suggestions are offered to improve the practical value of preservation policies by aligning them to the preservation maturity levels as developed by Dollar and Ashley [18].

2. GUIDANCE ON PRESERVATION POLICIES

2.1 Preservation standards about policies

In order to get an answer on the question: "what is a good preservation policy?" two standards are relevant for the preservation community. The OAIS model and the TDR standard.

The OAIS standard is a widely accepted standard in this community and offers a shared language for all practitioners. Although the exact phrase "preservation policy" is not mentioned in the OAIS standard, frequent references are made to "a policy" that an organization needs to formulate, related to several topics. These topics include a "pricing policy", a policy covering the deletion of objects and "access policies". In the OAIS standard, there are no prescriptions given for the elements that should be part of such a policy. It is left to the digital archive to decide what to include in which kind of policy. The entity "Management", as described in the OAIS functional model, is supposed to manage these policies, in relation to the broader policy domain in which the organization operates.

After the first publication of the OAIS model in 2002, the rather abstract concepts were further explained and described in "Trustworthy Repositories Audit & Certification (TRAC): Criteria and Checklist", published in 2007 [48]. TRAC states that an organization "has publicly accessible definitions and policies in place to dictate how its preservation service requirements will be met"[48], metric A3.1]. But no list of topics that should be included in policies was given.
This document was updated and augmented, and fairly recently resulted in the ISO standard 16363: Audit and Certification of Trustworthy Digital Repositories (TDR), finalized in 2012. This standard describes the criteria on which a certification of a digital archive will be based, explained in "metrics". As part of the "Policy Framework" the term "Preservation Policy" was introduced. The context for the Preservation Policy is explained as follows:

> *A repository is assumed to have an overall Repository Mission Statement, part of which will be concerned with preservation. The Preservation Strategic Plan states how the mission will be achieved, in general terms with goals and objectives. The Preservation Policy then declares the range of approaches that the repository will employ to ensure preservation (that is, to implement the Preservation Strategic Plan), and finally the Preservation Implementation Plan translates those into services that the repository must carry out [48], p. 1-4]* .

This policy framework is an abstract model and not prescriptive, in practice it might result in different documents under different names. Nevertheless a requirement of the Preservation Policy is that

> *The policies should be understandable by the repository staff in order for them to carry out their work. Preservation Policies and procedures must be demonstrated to be understandable and implementable. [48], p. 1-4*

So in order to be useful for staff, preservation policies should be realistic, otherwise they are not "understandable", let alone "implementable". The TDR standard does not give an overview of elements that should be covered in the Preservation Policy. However, in several ‘metrics’ that are outlined in the standard, references are made to preservation policies that should be in place, in order to show evidence of meeting the requirements mentioned in the metric. From this we can derive that preservation policies should be formulated addressing (at least) the following areas:
The existence of policies (3.3.2), periodic review of the policies (3.3.2.1), handling of liabilities, ownership and rights (3.5.1.4.), information integrity measures (3.3.5), a description of the collection that an organization "will preserve, retain, manage and provide access to" (3.1.3), the intellectual property rights (3.5.2), verification of the SIP on completeness and correctness (4.1.5), specifying the treatment of AIPs and the circumstances under which AIPs will be deleted (4.4.1.1), the properties to preserve (4.1.1.1, 4.1.1.2), preservation strategy and triggers to activate the strategy (4.3.1), changes in the preservation plan as a result of monitoring activities (4.3.3.), monitoring and acting upon hardware and software changes (5.1.1.1, 5.1.1.1.2, 5.1.2).
While both OAIS as well as TDR emphasize the importance of policies and TDR gives a clear definition of the need for such a policy, there is no overview of the elements that should be part of a preservation policy. This lack of guidance will not make it easy for organizations to develop a preservation policy. Lack of a shared understanding of the ingredients of a preservation policy will also make it difficult to judge, for example in an audit and certification procedure, whether a preservation policy is meeting the requirements or to compare archives.

2.2 Sources and guidance

In several published preservation policies, which will be discussed later, references were made to literature that supported and inspired the creation of these preservation policies. Taking policies of peer organizations is one way to get inspired in writing your own policies and several times the policies of the National Library of Australia [26] were used. Other sources included the JISC publication of Digital Preservation Policies Study by Neil Beagrie et all [4], and the Erpanet Tool [17]. What are these sources telling us about preservation policies?

The National Library of Australia, according to their website, created their first preservation policy in 2001 and in combination with their reputation in digital preservation, this might be a good reason why other organizations referred to these policies (the current version of their preservation policy is 0.4 and the 2001 version is available via their Pandora web archive [27]).

The Erpanet Digital Preservation Policy Tool (2003) starts with giving a set of general principles [17], p.3] for creating a policy, which "needs to convey the very philosophy of an organization concerning digital preservation". Another principle being: "every policy should be practicable, not definitive, capable of being put into practice by institutions with varying resources and needs, and, especially, flexible to adapt itself to changing administrative and technological circumstances". The Policy Tool further lists benefits of a preservation policy and offers an overview of elements that should be described in the preservation policy.

The motivation for the JISC publication lies in the evolving world of e-infrastructure and electronic services in universities and colleges in the UK, being dependent for future benefits on "digital preservation strategies being in place and underpinned by relevant policy and procedures" [4], p.5] . The JISC Digital Preservation Policies Study, published in 2008 provides a model for "Institutional Digital Preservation Policies" and is a practical guide with a set of policy clauses, further explained with examples in the implementation section. In Part 2 of this study several real life strategies of UK universities were analyzed. Areas where preservation policies could contribute to the mission and strategies of the universities were identified, thus demonstrating the need for consistency in preservation policies.

These sources offer guidance in creating a preservation policy and the topics that should be covered, at a time where OAIS and TRAC were fairly recently published and TDR was not yet a formal standard. These documents, together with the aforementioned standards were input for the creation of the SCAPE Catalogue of Policy Elements.

3. THE SCAPE PROJECT

3.1 The SCAPE Policy Framework

The European SCAPE project (running from 2010-2014) is dedicated to the digital preservation challenges of large scale, heterogeneous collections of complex digital objects. The project focuses on digital objects held in the collections of various participating content holders, with a focus on libraries, web archives and data centers. The scale of these digital collections limits the possibility of manual involvement when performing preservation activities. Instead, such large scale collections will require more automation through the use of workflows and high-performance systems. As preservation activities need to be guided by preservation policies, these policies will need to be formulated in such a way that they are machine readable (e.g. right level of granularity), in order to be usable in such automated processes. The focus of the policy work in the SCAPE project was on the activities for Preservation Watch and Preservation Planning. Starting point here was the Planets Functional View [37], where in addition to the OAIS model, the Preservation Watch function was added to the OAIS Functional Model, combining several monitoring functions. A Preservation Watch Function will need preservation policies in order to monitor the relevant areas and to determine these areas. In addition, the Preservation Planning will need these preservation policies to make a relevant plan.

The SCAPE Policy Framework (see Figure 1) was developed during the project, consisting of three levels:

1. Guidance Policies, a high level representation that describes in a broad sense the goals of the organization in relation to long term preservation of their collections. These Guidance policies can be derived from a strategy document or the mission of an organization.

2. Preservation Procedure Policies in which the approach to be taken to achieve the high level goals is described.

3. Control policies. On this level the policies formulate the requirements for a specific collection, a specific preservation action and/or for a specific designated community. This level can be human readable, but should also be machine readable and thus can be used in automated planning and watch tools to ensure that the chosen preservation actions and workflows meet the specific requirements identified for that digital collection. These are likely to be kept internally within the organization.

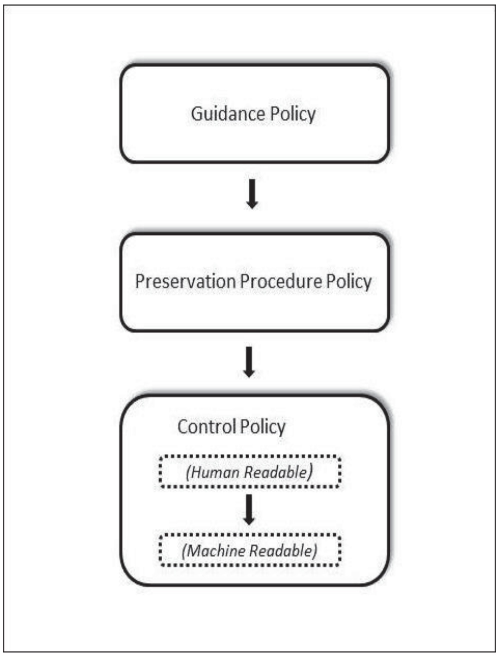


Figure 1 The SCAPE Policy Framework

Based on this framework, organizations should be able to create a set of policies that is consistent. This concept is described in [39] and presented during iPRES 2013.

The SCAPE framework can be mapped to the concepts described in the TDR standard. The Preservation Strategic Plan can be compared with the concept of Guidance Polices. The Preservation Policy has similarities with as the Preservation Procedure Policies and the Preservation Implementation Plan will result in the Control Policies.

3.2 The SCAPE Catalogue of Policy Elements

The previously mentioned standards and guidelines on preservation policies, as well as several other sources, were input for a set of topics that should be described on a strategic level. These topics were the basis for the Guidance Policy. Examples of these topics are: the use of a reference model for digital preservation, the concept of authenticity, whether the organization will preserve the digital material on bit preservation level or functional preservation level, whether access to the digital collection will be given, a view on the use of standards, the handling of various rights, etc.

This set of high level topics was input for the development of the SCAPE Catalogue of Policy Elements [9] in which the second level, the Preservation Procedure Policies, are described. Each policy element is described on the basis of a template. In this template the details of each policy element will give information about the need for the specific policy element and the risk of not having such a policy, the relationship with the strategic level, as well as the relationship with the lowest level, the Control Policy. A suggestion is made for the stage in the DCC Preservation Life Cycle [43] in which the policy will be created and who in the organization could be responsible for the description of the policy. These suggestions intend to connect the policy to the daily environment in which it should operate. Whenever relevant, elements are illustrated with an example of a real life policy.

This SCAPE Catalogue of Policy Elements is publicly available, both as a report as well as a wiki.

3.3 Published Preservation Policies

In several surveys [7], [44] organizations indicate that they have preservation policies in place. Some of them have published these policies on their website. To support the policy work in the SCAPE project a collection of published preservation policies was created and made publicly available on the website of the Open Planets Foundation [35]. In March, 2014 this set contained around 50 policies of libraries, archives, data centers and other organizations. The collection is created using a range of sources, including literature references, Internet search findings, direct contacts, responses to a blog post [38] that did an appeal on the digital preservation community to send references to publicly available preservation policies, suggestions from network partners and last but not least the incorporation of sources mentioned in the report published by Sheldon [36] in 2013, who did a similar exercise in collecting preservation policies.

The result is a highly heterogeneous set of preservation policies, from a large variety of organizations. When available, both the strategy as well as the policy are included in the collection. The boundaries between a preservation strategy and a preservation policy are not always clear, for example in [6]. In this example, many detailed approaches about the implementation of the strategy are described in the strategy, which other organizations are likely to describe under the heading "policy".

This initial collection of preservation policies cannot be seen as representative. Several organizations that are active in digital preservation are missing, either because they did not make their preservation policy publicly available, or because we did not find it — but that does not imply that they have no policy. As said before, all organizations with a preservation mandate will take decisions about their digital collections that are implicitly based on policies, whether they are written down or not. Each activity, whether it is the design or the selection of a preservation system, the operation of it, the planning of ingest procedures, or staff training, has its foundation in a vision on how to preserve the digital material.

By putting this overview of preservation policies on the Open Planets Foundation website, a central place is created where every organization, planning to develop or updating its preservation policies, can have a look at the policies of their colleagues and add their policies as well.

3.4 Analysis and observations

As already mentioned, this collection of preservation policies was originally created to validate the elements in the Catalogue of Preservation Policies. To support this validation a subset of around 40 published preservation policies was created. This selection was based on categories libraries, archives and data centers, as these organizations corresponded to the organizations that were the focus domains in the SCAPE project (web archives, digital repositories and data centers). All included policies were either in English or German.

Table 1: overview of analyzed preservation policies

Libraries	Archives	Data Centers
[3] ,[5] ,[6] ,[12] ,[15] ,[16], [22], [23] , [26] , [28] ,[29] , [31], [33], [40], [41], [46], [47], [50], [53], [54], [55], [56], [58]	[11], [13], [21], [24], [25], [30], [42], [32], [45], [57]	[10], [14], [19], [20], [49], [51], [52]

[numbers referring to references]

The main finding from the analysis was that almost all elements in the SCAPE catalogue were mentioned in the various policies, sometimes briefly and sometimes more extensively. Although the coverage of the SCAPE Catalogue of Policy Elements is broad, based on this analysis some elements seemed to be missing in the Catalogue. One reason for this, was that these elements were often not related to Preservation Watch and Preservation Planning (the focus areas), but to general aspects of policies. In a few cases there were elements in the policies that were also advised to include by the JISC report or the Erpanet Tool. Some examples of these elements are:

- A description of the review schema for the policies
- The explicit intention of the organization to collaborate with members from the digital preservation community, be it on the basis of knowledge exchange, contributing to standards, advising producers of digital material (especially in the policies of archives) or to be part of a network of digital archives
- A description of challenges that the organization is facing. Often a list of threats is given that the organization is facing with this mandate of digital preservation, like the rapid growth of digital material, the technical developments, sustainability, content provider partnerships, the needed flexibility, etc.
- Two preservation policies had a statement on the explicit intention to do research on digital preservation [30],[45]]
- As a preservation policy often has connections with other policies that are used in the organization, references are made to other relevant policies.

A general observation is that the 40 preservation policies differ greatly from one another. While sharing the same heading of "Preservation Policy", a highly heterogeneous set of documents was published. Although many of the elements from the SCAPE Catalogue were present in the policies, the level of detail used to describe these elements differs significantly. To give an indication of these different levels of detail: the length of the preservation policies ranges between 2 to 20 pages. While some were published in 2007 and have not been updated since, others were published as recently as 2014 [14]. In some cases, the policy is more or less a description of how digital preservation in general "should be" done, while others have very detailed descriptions of how this organization implemented various aspects of digital preservation [21]. A few try to combine their preservation policy for both analogue and digital material in one document [28],[56] but these are exceptions. Almost all of them focus only on the digital collection, as often also mentioned in the title of the document (e.g. "Digital preservation policy"). The content of the digital archive and the kind of collections that will be preserved, is often described in broad terms, as to explain which material will be affected by the policies described, using phrases like "digital born" material, digitized material (sometimes called "surrogates" [45]) and digital material on physical carriers. Subscriptions or licensed materials are often excluded from digital preservation [54]. In a few cases the preservation policy also contained the digitization policy, with sometimes detailed descriptions. [45] This was perhaps included because the digitization policy could lead to more digital material from their current analogue collections. The appendices have a variety of material, ranging from a list of supported formats, to a digital preservation decision flowchart [55] on the basis of which it is decided whether the digital object will be bit preserved or will get a full preservation treatment or not archived at all.

As mentioned before, the boundaries between a "policy" and a "strategy" are not entirely clear. For example, one organization wrote "This policy outlines the Record Office's approach to digital preservation, whilst the aim of the strategy is to describe this approach in more detail, including technical specifications where appropriate" [21], while others describe their approach in detail in the policies [31].

Although the standards mentioned earlier in this paper are clear about the need for a preservation policy, a set of criteria of what makes a good preservation policy is lacking. Several of the policies were written before TDR became a standard. This lack of criteria obviously lead to the heterogeneity in the set of existing preservation policies we found and one could wonder whether these preservation policies are playing the role they are intended to play. According to TDR the purpose of a preservation policy can be seen as "*declares the range of approaches that the repository will employ to ensure preservation"*. Staff will use the information in the preservation policy to shape their daily work in preserving digital material. Preservation policies should be clear enough to support this role. If we agree on the need for transparency, not only to staff but also to a broader range of stakeholders, the preservation policies have a role informing this - not specifically mentioned- audience, namely the audience that is interested in the trustworthiness of the repository. This does not imply that all policies should be publicly available, one could expect that different versions will exist to inform different stakeholders.

In order to fulfill these two roles, it should be clear what should be the essential elements in a preservation policy. On top of that, the need for clear criteria to assess the Policy Framework, including the Preservation Policies of an organization, will be necessary for the certification process based on TDR.

Looking at the heterogeneity in the current set of preservation policies, one could doubt very much whether all preservation policies are playing the role they were intended to do. And whether both the external stakeholders as well as the internal staff do see them as informing them adequately about the preservation approaches of the organization. In several cases this is debatable, for example if a policy promised a regular 2 year update and has not been updated for several years. As the digital preservation environment is changing, does not this show that the published document has not been adapted to this changing environment? And if not adapted to changes, what role does this policy play for the daily activities of the staff ?

But perhaps the current policies did not have these purposes and roles in mind when they were written? Why were they written and published anyway and for whom were they written, who are the stakeholders that needed to be informed? Depending on audience ("who") and purpose ("why"), different elements might have been described. What are the published preservation policies telling us about the "who" and "why"?

In general the "why" is more often addressed than the "for whom". Several reasons are given for why these policies are published, including "to state and communicate the principles" [24], "to describe the need and strategies for preserving (…) resources" [15], "[to] be an external statement of the current understanding and vision of digital preservation" [1], "[for] transparency" [31], "[to] define the principles" [40], "to formalize its commitment" [54], "to provide a comprehensive statement on the preservation and conservation of the Library's collections" [56] or "to outline what we can hope to achieve in the way of preserving digital material" [11]. In some cases this reason is related to the intended audience, but more often the audience is not mentioned at all. If it is mentioned, it is sometimes related to staff, but more often to what we can see as producers, consumers and funders. Two cases explicitly mention "peers (for general international evaluation of policies in this special area)" [46] and "the interested public as well as expert circles in the digital archiving /community" [42]. Based on these statements, it seems that the role of these policies are more focused on telling the intended public about their commitment in general. They seem to be less intended as guiding the daily practices in digital preservation, and staff activities. This could explain why in some policies firm commitments are phrased, while it is general knowledge in the digital preservation community that these commitments cannot be met yet. For example, a statement that all content will be validated on ingest is not a realistic one, as there are no tools for all file formats to do this. Similarly, a statement saying that a preservation strategy like emulation will be chosen is somewhat unrealistic if it is not really implemented and the current situation is that very few organizations have done so.

So how realistic are the published policies? In certification terminology: how trustworthy are the policies in giving evidence about what is really going on in an organization? Do they cover the previously given TDR description: "*The policies should be understandable by the repository staff in order for them to carry out their work. Preservation Policies and procedures must be demonstrated to be understandable and implementable*."? Apart from this TDR description, the problem of unrealistic phrases in preservation policies, is that this could be misleading both staff, the external stakeholders and the general audience. Digital preservation is an evolving field. Promising more in preservation policies then can be realized in practice could be dangerous and could lead to undermining the trustworthiness of the digital preservation community as a whole.

Adapting the preservation policies into documents that better reflect the real situation in an organization, should be a collaborative approach. As there is a variety of organizations with a long term preservation mandate, it might be useful to link a preservation policy to a preservation maturity level. If we could come to a set of criteria for each preservation maturity level, the need to promise more than the organization is capable of in that stage will be less. But these criteria should be based on a shared view in the digital preservation community. As preservation policies need to be updated on a regular basis, the organization can adapt the preservation policies to the next maturity level in a newer version and aligning them to the developments in the digital preservation community. The intentions of the organizations will be described in the Strategy, so it will be clear to the stakeholders and staff what are the goals and missions for the long term.

3.5 Levels of Maturity

Digital Preservation in an organization not seldom follows a long process before it is part of the daily practice. It could start with a project to get acquainted with the challenges and risks, followed by a more formal approach and finally ending up in making digital preservation an integral part of the organizations activities. Each phase will require a different strategy and a different set of preservation policies. Assigning a maturity level to an organization might give more insight in which stage of the process an organization is. One example of maturity levels is the Digital Preservation Maturity Level Model, developed by Dollar and Ashley [18] . This model distinguishes 5 stages. These stages are cited here, where for each stage a suggestion is made of the completeness of the preservation policy.

- Level 0, described as "Most, if not all, electronic records that merit long-term are at risk". In this situation an organisation is still figuring out what to do and a preservation policy is not yet expected.

- Level 1 "Many electronic records that merit long-term preservation are at risk.". The organization is more aware about its digital collection and might have started with approaches to handle these. A preservation policy will be under development, whereby projects done on part of the collections will give input.

- Level 2 "In this environment some electronic records that merit long-term preservation remain at risk". This would imply that the majority of the electronic records are taken care of and that only part of the overall digital collection is at risk. That is the moment a digital preservation policy should be in place and widely disseminated amongst the stakeholders, who should be convinced that the organization is knowing what it is doing. The preservation policy should be detailed enough for staff to develop procedures, related to the selected collections preserved.

- Level 3 "Few electronic records that merit long-term preservation are at risk". The existing preservation policy is regularly reviewed an updated and will be detailed enough for staff to develop procedures for all collections preserved.

- Level 4 "no electronic records that merit long-term preservation are at risk". The preservation policy is continuously reviewed and updated as needed, based on developments.

Although this linking of maturity level and the state of the preservation policy can be a step forward, it does not solve the problem that some organizations will describe approaches in their policies they are not experiencing yet. This could be solved if suggestions for a minimal set of elements per policy on each level would be added to this maturity level model. For example, on level 1 one could expect that organizations are doing bit preservation, so they do not need to cover migration or emulation in their policies (but they might cover this in their strategy), as they most likely will not use these preservation approaches yet. However, they do need to mention metadata and the use of standards in more detail, as these are elements that influence the design of the AIP. A broadly accepted model in which for every maturity level is explained what should be the essential elements in a preservation policy, will support organizations to come to more realistic preservation policies. But how do we get there?

4. FURTHER ACTIONS

The TDR ISO 16363 standard offers a clear description of the role a preservation policy is supposed to play in a digital repository. But an analysis of published preservation policies shows that there is a gap between the expectations in the standards and how the policies are stated in practice. Preservation policies that are promising more than can be realized are a risk for both the trustworthiness of the organization as well as for the preservation community.

Although the work on policies in SCAPE is finished, the further development of the SCAPE Catalogue of Policy Elements will be sustained after September 2014 as a wiki on the website of the Open Planets Foundation, publicly available. Everyone in the digital preservation community can suggest new elements to be added, according to the template that has been created. This Catalogue of Policy Elements offers a good starting point to develop guidelines for creating preservation policies, more adapted to the maturity level of an organization, leading to more realistic preservation policies. These cannot only be used to inform staff and other stakeholders but will be valuable in the audit and certification process to assess the Policy Frameworks of organizations. Such Policy Frameworks will also contribute to the trustworthiness of the digital preservation community in general.

5. ACKNOWLEDGMENTS

My thanks to Catherine Jones (Science and Technology Facilities Council, UK), Gry Elstroem (State and University Library Aarhus, Denmark) and Sean Bechhofer (University of Manchester, UK) for the inspiring collaboration on the policy work in the SCAPE project.

This work was partially supported by the SCAPE Project. The SCAPE project is co-funded by the European Union under FP7 ICT-2009.4.1 (Grant Agreement number 270137).

6. REFERENCES

[1] Archives New Zealand and National Library of New Zealand: Digital Preservation Strategy, 2011. Retrieved 22-03-2014 from http://archives.govt.nz/sites/default/files/Digital_Preservation_Strategy.pdf

[2] Audit and Certification of Trustworthy Digital Repositories, 2012 http://public.ccsds.org/publications/archive/652x0m1.pdf

[3] Bavarian State Library: Digital Preservation Policy, 2012. Retrieved 22-03-2014 from http://www.babs-muenchen.de/content/dokumente/2012-11-22_BSB_Preservation_Policy.pdf

[4] Beagrie, N, Semple,N, Williams, P, Wright, R: Digital Preservation Policies Study, 2008 Retrieved 22-03-2014 from http://www.jisc.ac.uk/media/documents/programmes/preservation/jiscpolicy_p1finalreport.pdf

[5] Boston University Library: Digital Preservation Policy. Retrieved 22-03-2014 from http://www.bu.edu/dioa/openbu/boston-university-libraries-digital-preservation-policy/

[6] British Library: Digital Preservation Strategy 2013-2016. Retrieved 22-03-2014 from http://www.bl.uk/aboutus/stratpolprog/collectioncare/discovermore/digitalpreservation/strategy/BL_DigitalPreservationStrategy_2013-16-external.pdf

[7] Canadian Heritage Information Network survey 2011 Retrieved 22-03-2014 from: http://bit.ly/16HS7Cj

[8] Candela, L. and A.Nardi (ed.) *Digital Library Technology and Methodology Cookbook* Retrieved 24-04-2013 from http://www.dlorg.eu/index.php/publications

[9] Catalogue of Preservation Policy Elements, SCAPE project http://www.scape-project.eu/wp-content/uploads/2014/03/SCAPE_D13.2_KB_V1.0.pdf

[10] CenterData: Preservation and Dissemination Policy of the LISS Data Archive. Retrieved 22-03-2014 from http://www.lissdata.nl/assets/uploaded/reservation%20and%20Dissemination%20Policy%20of%20the%20LISS%20Data%20Archive_1_0.pdf

[11] Cheshire Archives: Digital Preservation Policy, 2010 Retrieved 22-03-2014 from http://archives.cheshire.gov.uk/record_care/digital_preservation/digital_preservation_policy.aspx

[12] Cornell University Library: Cornell University Library Digital Preservation Policy Framework. Retrieved 22-03-2014 from http://ecommons.library.cornell.edu/handle/1813/11230

[13] Danish National Archives: *Digital Preservation Policies* Retrieved 22-03-2014 from http://www.sa.dk/media%284826,1033%29/Strategy_for_archiving_digital_records.pdf

[14] DANS (Data Archive and Networked Services),Preservation Policy, 2014. Retrieved 22-03-2014 from http://dans.knaw.nl/sites/default/files/file/EASY/20140220%20Preservation%20Policy%20v1_0.pdf

[15] Dartmouth College Library: Digital Preservation Policy. Retrieved 22-03-2014 from http://www.dartmouth.edu/~library/digital/about/policies/preservation.html?mswitch-redir=classic

[16] Deutsche National Bibliothek: Langzeitarchivierungs-Policy der Deutschen Nationalbibliothek. Retrieved 22-03-2014 from http://nbn-resolving.de/urn:nbn:de:101-2013021901

[17] Digital Preservation Policy Tool http://www.erpanet.org/guidance/docs/ERPANETPolicyTool.pdf

[18] Dollar, C. & Ashley, L.: Assessing Digital Preservation Capability using a Maturity Model Process Improvement Approach.2013 Retrieved on 20-03-2014 from https://docs.google.com/file/d/0BwbqtwrvKHokR3g0RVR4bmNIWjg/edit?usp=sharing

[19] Florida Digital Archive. FDA Policy and Procedures (2011) . Retrieved 22-03-2014 from https://fclaweb.fcla.edu/uploads/FDAPolicyGuideversion3.0.pdf

[20] GESIS Data Archive for the Social Sciences. *Digital Preservation Policy (2014). Retrieved 22-03-2014 from* http://www.gesis.org/fileadmin/upload/institut/wiss_arbeitsbereiche/datenarchiv_analyse/DAS_Preservation_Policy_eng.pdf

[21] Hampshire Archives at Hampshire Record Office : Digital Preservation Policy, 2010. Retrieved 22-03-2014 from http://www3.hants.gov.uk/archives/hro-policies/hro-digital-preservation-policy.htm

[22] Hathi Trust Digital Library: Digital Preservation Policy. Retrieved 22-03-2014 from http://www.hathitrust.org/preservation

[23] John Hopkins Sheridan Libraries: Digital Preservation Policy. Retrieved 22-03-2014 from http://old.library.jhu.edu/collections/institutionalrepository/irpreservationpolicy.html

[24] London Metropolitan Archives: Interim Digital Preservation Policy. 2010. Retrieve 22-03-2014 from http://217.154.230.218/NR/rdonlyres/6466F6FA-2F04-4E3E-8D8D-9158FD303425/0/DigitalPreservationPolicyJun2010.pdf

[25] National Archives of Australia. *Digital Preservation Policy. 2011. Retrieved 22-03-2014 from* http://www.naa.gov.au/about-us/organisation/accountability/operations-and-preservation/digital-preservation-policy.aspx

[26] National Library of Australia Digital Preservation Policy 4th Edition. Retrieved on 20-03-2014 from http://www.nla.gov.au/policy-and-planning/digital-preservation-policy

[27] National Library of Australia: A Digital Preservation Policy for the National Library of Australia, 2001.Retrieved 21-03-2014 at http://nla.gov.au/nla.arc-36099

[28] National Library of Finland: Preservation Policy Retrieved 22-03-2014 from http://www.kansalliskirjasto.fi/attachments/5v5daJ8e3/5pzFQo6pJ/Files/CurrentFile/NLF_Preservation_Policy.pdf

[29] National Library of Wales: Digital Preservation Policy and Strategy. Retrieved 22-03-2014 from http://www.llgc.org.uk/fileadmin/documents/pdf/2008_digipres.pdf

[30] Parliamentary Archives: A digital preservation policy for Parliament. 2009 Retrieved 22-03-2014 from http://www.nationalarchives.gov.uk/documents/tna-corporate-preservation-policy-2009-website-version.pdf

[31] Portico,Portico Preservation Policies, 2013 Retrieved 22-03-2014 from http://www.portico.org/digital-preservation/about-us/portico-resource

[32] Public Record Office of Northern Ireland: Digital Preservation Strategy. 2013 Retrieved 22-03-2014 from http://www.proni.gov.uk/digital_preservation_strategy.pdf

[33] Purdue University Research Repository: Digital Preservation Policy. Retrieved 22-03-2014 from https://purr.purdue.edu/legal/digitalpreservation

[34] Reference Model for an Open Archival Information System, 2012 CCSDS retrieved on 20-03-2014 from http://public.ccsds.org/publications/archive/650x0m2.pdf

[35] SCAPE Published Preservation Policies. Retrieved on 22-03-2014 from http://wiki.opf-labs.org/display/SP/Published+Preservation+Policies

[36] Sheldon, M Analysis of current digital preservation policies archives, libraries, and museums.2013 Retrieved on 20-03-2014 from

http://www.digitalpreservation.gov/documents/Analysis%20of%20Current%20Digital%20Preservation%20Policies.pdf?loclr=blogsig

[37] Sierman, B. and Wheatly, P: Evaluation of Preservation Planning within OAIS, based on the Planets Functional Model. Planets Project 2010 Retrieved 22-4-2013 from http://www.planets-project.eu/docs/reports/Planets_PP7-D6_EvaluationOfPPWithinOAIS.pdf

[38] Sierman, B: Published Preservation Policies. 7-10-2013 Blogpost on Open Planets Foundation, retrieved 22-03-2014 http://www.openplanetsfoundation.org/blogs/2013-10-07-published-preservation-policies

[39] Sierman, B Jones, C Elstroem, G, Bechhofer S: Preservation Policy Levels in SCAPE. Ipres 2013 retrieved 20-03-2014 from http://purl.pt/24107/1/iPres2013_PDF/Preservation%20Policy%20Levels%20in%20SCAPE.pdf

[40] State Library of Queensland: Digital Preservation Policy, 2008. Retrieved 22-03-2014 from http://www.slq.qld.gov.au/__data/assets/pdf_file/0020/109550/SLQ_-_Digital_Preservation_Policy_v0.05_-_Oct_2008.pdf

[41] Statsbiblioteket State and University Library, Denmark Digital Preservation Policy for State and University Library Denmark, version 2.0, 2012 Retrieved 22-03-2014 from http://en.statsbiblioteket.dk/about-the-library/ddpolicy

[42] Swiss Federal Archive: Digital Archiving policy, 2009. Retrieved 22-03-2014 from http://www.bar.admin.ch/themen/00876/index.html?lang=en&download=NHzLpZeg7t,lnp6I0NTU042l2Z6ln1ad1IZn4Z2qZpnO2Yuq2Z6gpJCDdYB,fmym162epYbg2c_JjKbNoKSn6A--

[43] The DCC Curation Life cycle model. Retrieved 22-03-2014 from http://www.dcc.ac.uk/sites/default/files/documents/publications/DCCLifecycle.pdf

[44] The Digital Divide.Assessing Organisations'Preparations for Digital Preservation. Retrieved 22-03-2014 at http://www.planets-project.eu/docs/reports/planets-market-survey-white-paper.pdf

[45] The National Archives: Preservation Policy. 2009. Retrieved 22-03-2014 from http://www.nationalarchives.gov.uk/documents/tna-corporate-preservation-policy-2009-website-version.pdf

[46] The Royal Library: The National Library of Denmark and Copenhagen University Library: Policy for long term preservation of digital materials at the Royal Library, 2010. Retrieve 22-03-2014 from http://www.kb.dk/export/sites/kb_dk/da/kb/downloadfiler/PreservationPolicyDigitalMaterials_21092012.pdf

[47] The University of Manchester Library: Digital Preservation Strategy (2012). Retrieved 22-03-2014 from http://www.library.manchester.ac.uk/aboutus/strategy/_files2/Digital-Preservation-Strategy.pdf

[48] Trustworthy Repositories Audit & Certification:Criteria and Checklist, 2007 http://www.crl.edu/sites/default/files/attachments/pages/trac_0.pdf

[49] UK Data Service: Preservation Policy (2012). Retrieved 22-03-2014 from http://data-archive.ac.uk/media/54776/ukda062-dps-preservationpolicy.pdf

[50] University of Massachusetts Amhurst Libraries: Digital Preservation Policy.2011 Retrieved 22-03-2014 from http://www.library.umass.edu/assets/aboutus/attachments/University-of-Massachusetts-Amherst-Libraries-Digital-Preservation-Policy3-18-2011-templated.pdf

[51] University of Illinois at Urbana-Champaign: IDEALS Digital Preservation Policy, 2009 Retrieved 22-03-2014 from https://www.ideals.illinois.edu/bitstream/handle/2142/2383/IDEALS_PreservationPolicy_Nov2009.pdf?sequence=4

[52] University of Michigan: ICPSR Digital Preservation Policy Framework. Retrieved 22-03-2014 from http://www.icpsr.umich.edu/icpsrweb/content/datamanagement/preservation/policies/dpp-framework.html

[53] University of Minnesota: University Digital Conservancy Preservation Policy. Retrieved 22-03-2014 from http://conservancy.umn.edu/pol-preservation.jsp

[54] University of South Carolina Libraries: Digital Preservation Policy Framework.2010. Retrieved 22-03-2014 from http://library.sc.edu/digital/USC_Libraries_Digital_Preserva.pdf

[55] University of Utah J. Willard Marriot Library: Digital Preservation Policy, 2012. Retrieved 22-03-2014 from http://www.lib.utah.edu/collections/digital/DigitalPreservationPolicy2012.docx

[56] Wellcome Library: Preservation Policy for Materials held in Collections. Retrieved 22-03-2014 from http://wellcomelibrary.org/content/documents/policy-documents/preservation-policy

[57] West Yorkshire Archive Services. Digital Archives Policy.2007 Retrieved 22-03-2014 from http://www.archives.wyjs.org.uk/documents/archives/WYAS%20Digital%20Archives%20Policy.pdf

[58] Yale University Library. Digital Preservation Policy. Retrieved 22-03-2014 from http://www.library.yale.edu/iac/DPC/revpolicy2-19-07.pdf

Integrating e-government systems with digital archives

Kuldar Aas
National Archives of Estonia
J. Liivi 4
Tartu, 50409, Estonia
+372 7387 543
Kuldar.Aas@ra.ee

Janet Delve
University of Portsmouth
Eldon Building
Winston Churchill Avenue
Portsmouth PO1 2DJ
+442392 845524
Janet.Delve@port.ac.uk

Ricardo Vieira
INESC/IST, Universidade de Lisboa
Rua Alves Redol nº 9
Lisbon, 1000-029, Portugal
+351 210 407 066
rjcv@ist.utl.pt

Ross King
Austrian Institute of Technology
Donau-City-Straße 1
Vienna 1220, Austria
+43 50550 4271
ross.king@ait.ac.at

ABSTRACT

While interoperability between active e-government systems has been a significant area of work during the last decade, the fact that much of this information needs to be preserved for the long-term after the initial creation has been ignored, and the re-use of data has been of secondary concern. This paper looks into the needs of long-term preservation of digital-born e-government data and describes how the EC-funded E-ARK project proposes further actions to address the challenge in a cost-effective manner.

General Terms

infrastructure, communities, strategic environment, digital preservation marketplace, case studies and best practice, training and education.

Keywords

long-term preservation, interoperability, digital repositories, OAIS, ingest, access, E-ARK, data mining, big data, e-infrastructures, e-government.

1. INTRODUCTION

The continued effort to develop and apply new and efficient e-government systems has created huge benefits in terms of back-office efficiency as well as in the way that citizens and businesses can interact with government institutions. However, until now most efforts have concentrated on the active phase of the information lifecycle (i.e. creation and short-term management of data) and little attention has been paid to the later stages (long-term preservation and access).

At the same time, national digital repositories (mainly national archives) have the mandate and obligation to ingest, preserve and offer long-term access to valuable pieces of government information irrespective of their format. While approaches to preserving analogue records have been long established, concomitant methods for handling digital-born information are lacking at present.

In order to fill this gap, the European Commission has funded a three year project called E-ARK[1] which includes a broad range of leading practitioners from all sides of the issue – records creators and e-government legislators, national archives, research institutes and software providers for both live data and digital preservation solutions.

2. CURRENT PRACTICES AND PROBLEMS

Multiple studies have been carried out recently to learn more about the maturity of digital preservation solutions. As an example the Danish [1], Belgian [2] and Swiss National Archives and projects like DC-NET [3], DCH-RP [4] and SCAPE [5] have provided studies on this topic.

While all of these studies concentrate on some specific issues in digital preservation, we can see that especially in the archival sector, practices for preserving digital information are just emerging, and there are only a few countries where digital preservation has indeed been applied in a practical and holistic way. In Europe most notably the UK, Danish, Norwegian, Swiss and Dutch national archives have established relevant procedures and systems to allow for the transfer, preservation and access of born-digital government data. Elsewhere also the US and Australian government and state sectors have been active and successful.

Based on the above mentioned studies, the E-ARK project has continued this work and carried out a comprehensive study in early 2014 to learn more about the technical details of the available solutions[2]. The results of the study show that even in the case where solutions are available, these are rather pragmatic approaches towards the generic problem in that they are limited to

[1] http://www.e-ark-project.eu/

[2] The results of the study are available in E-ARK deliverables D3.1, D4.1 and D5.1 here: http://www.eark-project.com/resources/project-deliverables

only addressing the immediate, pressing necessities of preservation and do not extend to re-use and access. In essence, typical national preservation requirements tend to consist of a set of metadata requirements which must be fulfilled by the agency transferring data, together with some formatting rules for the metadata as well as the actual data (i.e. regulations on archival file formats). The normal method of accessing preserved data is through archival catalogues, where users first face the burden of identifying relevant datasets before they are able to start looking for the bits of information they actually need.

The main reason for applying such fragmented approaches we have identified is the lack of standardization. While interoperability between e-government systems has been a major focus for governments across the world, not much has been done in harmonizing methods for data export and transfer to long-term storage facilities. The cumulative effect is that:

- each jurisdiction provides its own national standards for pre-ingest and ingest workflows as well as for the Submission Information Package (SIP) structure and content;
- in order to apply these standards, information systems' export functionalities involve custom development by all government institutions, thus making it a significant financial burden;
- the quality of data and, even more crucially, metadata harvested from source systems is often lacking due to the limited amount of resources the national archives have for developing relevant standards and offering training;
- the quality of data and metadata stored in digital repositories is often rather low and thus the information itself is hard to find for the persons actually needing access to it;
- due to problems with data quality, lack of common preservation data models, and lack of funding, it is hard for national archives to provide the kind of services expected by the general public, most notably the ability to offer preserved data for use by e-government infrastructures and central mash-up services in national and international service portals;
- as such the transfer of data to long-term storage includes a huge loss in accessibility, which makes the data owners less willing to undertake the actual transfer and more inclined to develop their own digital repositories, in turn spending a considerable amount of money to set up systems which do not constitute their core business and without possessing reasonable in-house knowledge of digital preservation.

3. NEED FOR STANDARDIZATION

From the discussion above it is clear that there is a need for standardizing key elements of the later phases of the information lifecycle. Special attention should be paid to the interoperability steps – actions during which data and metadata is transferred from one system to another, or accessed between these systems.

3.1 Export and transfer of born-digital data

In terms of standardising the export of data from source systems, there has already been some effort put into the creation of the MoReq2010 (Model Requirements for Records Systems) specification.[3] The MoReq2010 specification includes, among other parts, high-level requirements for the bulk export of records from systems for records archiving or system migration scenarios. However, the requirements are not sufficiently detailed to allow the development of interoperable technical components.

The goal of the E-ARK project is to build on the high-level MoReq2010 specification and update it by adding more detailed requirements derived from already available national best practices. In particular, the following elements must be available in order to support increased interoperability:

- a metadata schema that mandates the use of core elements for automating the export, validation and transfer workflows;
- extensibility options that would allow the addition of country or domain-specific metadata to the central core (as an example metadata specific to eHealth or eInvoice records);
- metadata re-use specifications to outline how metadata created in e-government systems might be re-used for archiving purposes;
- a pre-ingest and ingest workflow model that outlines the crucial actions of metadata, data integrity and authenticity validation;
- transfer mechanisms that allow the bulk transfer of records and their metadata from agencies to archives in an efficient and secure manner.

First drafts of these principles will be available within E-ARK as early as at the end of 2014.

Clearly the process of exporting records from source systems for the transfer to long-term repositories is conceptually not much different from exchanging information between any other systems. Therefore the task is being carried out in close cooperation with the EC-funded e-SENS (Electronic Simple European Networked Services) project[4], which aims to develop principles and specifications for cross-border services which exploit data from various European administrations.

The outcome of this task would in particular allow e-government projects and international software providers (like Oracle, Microsoft and others) to create native data export functionality that can easily be implemented across systems and jurisdictions. All of this will be possible at a fraction of the cost currently put into institutional or national custom developments.

The availability of common specifications will also allow the development of common training and dissemination programs, thus contributing to increased awareness in general as well as to the increase of data quality and understandability in archiving processes.

3.2 Open preservation formats

Another important aspect to be examined is the standardization and further development of Archival Information Package (AIP) structures and principles. During the first six months of the E-ARK project, a detailed analysis of current prevailing AIP principles has been carried out[5]. This analysis shows that there are already a good number of standards and specifications available as a starting point. For example, PREMIS[6] is widely used for

[3] http://moreq2010.eu/

[4] http://www.esens.eu/

[5] The analysis is available as part of the E-ARK deliverable D4.1 at http://www.eark-project.com/resources/project-deliverables

[6] http://www.loc.gov/standards/premis/

preservation metadata, EAD[7] for archival descriptions, the MoReq metadata module for records management descriptions and finally METS[8] and BagIt[9] for bringing all the different components together. Such an approach is especially visible in the recent AIP specifications provided by the Swedish National Archives[10] and the North-Rhine Westphalia state government[11].

The E-ARK project will continue to evaluate the already available standards and will define a limited core set of mandatory elements for all AIP packages. Most essentially, the elements which are needed for preservation planning, ensuring integrity and authenticity of archived data must be defined to allow for interoperability between different preservation systems.

In addition we aim to add support for additional access-oriented layers to the AIP specification:

- AIP Level 0: for structured data the Level 0 format will allow storage "as is" – with the original data model intact – while allowing for additional semantic enrichment of the contents as an OWL-oriented representation;
- AIP Level 1: the Level 1 AIP is created by analysing the Level 0 AIP and turning it into more easily usable OLAP (OnLine Analytical Processing[12]) cubes, following methods from data warehousing. As such, using a Level 1 AIP will allow the archives to offer easier access to data without the need to learn the specifics of the original data model;
- AIP Level 2: the Level 2 AIP is mainly intended to be used for archiving systems which hold unstructured records (as an example pdf files with common metadata).

These enhancements to currently available AIP formats as well as the tools which will be developed to support the formats, SIP to AIP conversion, and conversion between different AIP levels, will essentially allow archives to store in parallel the original database from which records originate as well as more user-friendly representations (as Level 1 or Level 2 representations) in a harmonised way.

In addition, the possibilities for semantic enrichment of content should be of interest for archives that preserve structured records. Namely, the use of semantic technologies will allow users to search for relevant data by using semantic entities instead of searching for relevant databases and useful elements in their data model. In other words, archives will be enabled to allow searching across database contents independent of their original data models.

3.3 Access to archived data

Currently most archives provide access to their digital holdings through dedicated archival catalogues. In addition, the archives tend to organize their content according to archival hierarchical classification schemes and description rules. This means that the "rich" metadata descriptions are usually available only for aggregations of data (collections) but not the single elements (i.e. records) which are mostly the scope of public interest.

To give an example, in most countries governmental information systems are currently being archived as database snapshots – the full content of a relational database created in a specified timeframe. This snapshot is usually migrated into open formats and the data model is technically described. At the same time, a content description is usually available only for the whole dataset. As a result, potential users interested in the information must first locate the relevant dataset(s) in the archives catalogue and then query all of these one by one. Added to that, changes in the functions of public sector agencies are also reflected in the scope of their information systems – in the long term the data on a specific topic might have moved between agencies and systems – thus making even the discovery of all relevant snapshots a difficult task.

The E-ARK project is working on a series of solutions in order to overcome these issues. The first approach involves the use of semantic description and data warehousing techniques. The key idea is that if all archived databases were to follow a single formatting specification, it would become possible to apply semantic description and data de-normalisation methods taken from data warehousing approaches [6]. As a result the entry point to preserved data would be simple semantically enriched OLAP cubes instead of relational database snapshots with highly complex data structures. This would allow users to browse the preserved data more easily as well as open new possibilities for data mining on top of the data preserved in the archives.

The other access method which is being researched inside E-ARK is the access to archived records from external systems (as an example government service portals or agency web sites). Again, when we can assume that all government records have been described in digital repositories by using common core metadata elements, it is fairly straightforward to produce API specifications for querying and accessing these records. In more detail, the project is looking at the OASIS standard CMIS (Content Management Interoperability Services)[13]. While CMIS describes the full range of CRUD services (Create-Read-Update-Delete), the application in a digital repository must limit the set to only Read and partial Update services. In addition, the workflow to negotiate for permanent ID (PID) creation and exchange must be examined as well as how to deal with active preservation methods where the technical characteristics of data can change over time (as an example, file format migration might have been applied).

Ultimately, the implementation of the "CMIS Application Profile for Archives" would allow institutions to transfer their data to digital repositories while still being able to continue offering the kind of data access services convenient for their users.

4. NEED FOR HARMONIZATION OF KNOWLEDGE

Despite the lack of standardization, information management (IM) has known extensive research and practice in the past years. In fact, nowadays many business and technical references have emerged to guide the processes of ingesting, managing, preserving and accessing information. In terms of designing the processes, standards such as ISO15489[14] ("Records Management") or

[7] http://www.loc.gov/ead/

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

[9] https://wiki.ucop.edu/display/Curation/BagIt

[10] http://riksarkivet.se/publicerade-rapporter-fran-eard (in Swedish)

[11] http://www.danrw.de/?lang=en

[12] http://searchdatamanagement.techtarget.com/definition/OLAP

[13] http://docs.oasis-open.org/cmis/CMIS/v1.0/cmis-spec-v1.0.html

[14] http://www.iso.org/iso/catalogue_detail?csnumber=31908

ISO30300/1[15] ("Management systems for records") already exist. For implementing tools and services references such as MoReq2010 and ISO16175[16] ("Principles and functional requirements for records in electronic office environments") are well known. For assessing organizations and tools one can refer to ISO16363[17] ("Audit and certification of trustworthy digital repositories") or ISO18128[18] ("Risk assessment for records processes and systems"). Also it is important to note that the examples above are international references. As noted in Section 3 above, due to lack of standardization, several countries have also been defining national practices and procedures that should not be discarded and constitute relevant knowledge to the field.

Apart from that, it is also important that, as a strategy, we admit we live in a world were problems can be seen from different perspectives and consequently solutions need to consider requirements from different areas. Applying this to IM, it means we must recognize other specific views besides only those of information science, such as also information systems, software engineering, risk management, among others. In fact, all this exists already, and is a concern usually known in the engineering and management areas as "Enterprise Architecture".

The proliferation of standards and references together with the recognition that problems should be analysed from different perspectives, motivations and communities has raised the need for a knowledge system that allows stakeholders to obtain a consolidated view of the existing knowledge. Therefore, one of the goals of the E-ARK project is to design and implement an online open-access knowledge centre that offers the possibility of uploading, managing and consolidating existing best practices, standards and other references not only in the core domain of IM but also in relevant peripheral domains. This service can then be used by business stakeholders in order to understand IM practices and requirements, IM stakeholders as a main source for information and knowledge, and/or, academics and students as a teaching and learning resource.

5. VALORIZATION OF ARCHIVAL DATA

As a final innovative aspect of the project, E-ARK will promote the re-use of archival data by facilitating a common pan-European approach to providing simple and advanced queries to researchers, to the public and private sectors, and to citizens. The project will research data mining techniques that will enable new forms of data re-use that provide vital decision support for business end users. This will in turn enhance competitive intelligence in the EC digital economy by providing analytical processing access to various longitudinal data sets: demographic, economic, judicial and so on. For example, data mining techniques could be used to compare house price fluctuations across various European cities over time to produce a vital pan-European economic dataset. This could be used as a basis for various marketing and research purposes.

At the other end of the spectrum, archival historical and cultural data could be marketed as part of commercial teaching packages. Data mining within E-ARK will allow analysis of aggregated sets of archival data, comprising structured and unstructured information, to identify new patterns of activity in consumer, business and systems behaviour. It should be possible to analyse open archive records to find trends, correlations, etc. It should be possible to apply post-ingest algorithms to archival records (e.g. automated classification based on machine learning). This ability to analyse activity, rather than survey, sample or observe, is transformational, in that genuine patterns of behaviour can be identified – thus providing a basis for new products and services. E-ARK will thus facilitate the taking up of opportunities afforded by "Big Data" that will become available as a result of archival interoperability.

6. SUMMARY AND TIMELINE

When considering current e-government solutions, we observe that in too many cases the long-term preservation approaches applied form a costly bottleneck in the holistic view of the data lifecycle management.

The E-ARK project plans to change this by providing a set of standardized specifications which can be implemented across borders and therefore allow all archival and other government institutions to apply, discuss and develop these further in a common and collaborative manner. E-ARK should also improve access to archived public information through standard query interfaces and data mining techniques.

While the E-ARK project is still in its early stages, the final results will be available and implemented in a reference implementation by early 2017 and first drafts of crucial elements (e.g. best practice reviews across the archival sector) will be available as early as autumn 2014.

7. ACKNOWLEDGMENTS

E-ARK is an EC-funded pilot action project in the Competitiveness and Innovation Programme 2007-2013, Grant Agreement no. 620998 under the Policy Support Programme.

8. REFERENCES

[1] Kristmar, K. V. (2012): "Common challenges, different strategies." Retrieved from http://www.sa.dk/media%284227,1033%29/1._Common_Challenges_-_Different_Strategies,_KVK.pdf

[2] Velle, K. (2012): "Database Archiving." Retrieved from https://www.sa.dk/media(4588,1033)/EBNA-Minutes,_CPH_29-30_May_2012.pdf

[3] Ruusalepp, R. & Dobreva, M. (2012): "Digital Preservation Services: State of the Art Analysis." Retrieved from www.dc-net.org/getFile.php?id=467

[4] Justrell B., Toller E. (2013): "Standards and interoperability best practice report." Retrieved from http://www.dch-rp.eu/getFile.php?id=165

[5] Faria L, Duretec K., Kulmukhametov A., Moldrup-Dalum P., Medjkoune L., Pop R., Barton S., Akbik A. (2014): "SCAPE survey on preservation monitoring." Retrieved rom http://www.scape-project.eu/wp-content/uploads/2014/05/SCAPE_D12.2_KEEPS_V1.0.pdf

[6] Inmon, W.H. Building the Data Warehouse, Fourth Edition, John Wiley and Sons, New York, 2005.

[15] http://www.iso.org/iso/catalogue_detail?csnumber=53732

[16] http://www.iso.org/iso/catalogue_detail.htm?csnumber=55790

[17] http://www.iso.org/iso/catalogue_detail.htm?csnumber=56510

[18] http://www.iso.org/iso/catalogue_detail.htm?csnumber=61521

Self-assessment of the Digital Repository at the State and University Library, Denmark - a Case Study

Gry Vindelev Elstrøm
State and University Library, Denmark
Aarhus, Denmark
Telephone number 0045 8946 2314
gve@statsbiblioteket.dk

Jette Gjelstrup Junge
State and University Library, Denmark
Aarhus, Denmark
Telephone number 0045 8946 2054
jgj@statsbiblioteket.dk

ABSTRACT

Organisations with a commitment to long-term digital preservation need to be perceived as trustworthy to meet the demands of their stakeholders. Audit and certification procedures provide a means to transparency and trustworthiness. The State and University Library has worked with trustworthiness for several years using different tools. In this paper, we describe the process and the benefits of performing an audit based on self-assessment by the use of ISO 16363 on the digital repository of the State and University Library. After describing the digital collections, DP organisation, policy framework and repository infrastructure it is explained how The State and University Library has been working with trustworthiness over the last four to five years. The latter part of the paper describes how we have conducted a self-assessment of the digital repository by the means of the ISO standard 16363. We explain some of the challenges of the work and the immediate effects of the process, which, at the time of writing, is not finished yet.

General Terms

Management, Measurement, Documentation, Reliability, Security, Standardization.

Keywords

ISO 16363, self-assessment, audit, metrics, digital preservation, metadata.

1. INTRODUCTION

In this paper, we describe the process, the benefits and the challenges of performing a self-assessment by the use of ISO 16363 [8] on the State and University Library's digital repository.

Over the last decade there has been an increasing interest among libraries and archives engaged with digital preservation to have their repository classified as trustworthy. This is also the case at the State and University Library, Aarhus, Denmark (hereafter referred to as "SB") where the work of becoming a trustworthy digital repository is seen as an on-going process, since there will always be room for improvements. Since 2010 a management team focusing on the library's digital collections has worked continuously with audit procedures in order to comply with audit criteria as part of the process of becoming a trustworthy digital repository.

The audit work is part of the library's strategy to enhance and develop its work on digital preservation. This is also in line with the library's national and international involvement in digital preservation initiatives, e.g. the Danish Net Archive (http://netarkivet.dk/), the Danish information site on digital preservation Digitalbevaring.dk (http://digitalbevaring.dk/), the EU-funded SCAPE project (http://www.scape-project.eu/) and Open Planets Foundation (http://openplanetsfoundation.com/).

Obliged by federal law SB preserves Danish cultural heritage in the form of large audio-visual collections of radio and television broadcasts, movie and TV commercials, sound recordings (voice and music), the Danish Net Archive, the Danish National Newspaper Collection etc. It is a broad and diverse span of collections with a large demand for control and curation to keep the collections preserved for the long term.

A self-assessment of the digital repository would expose all drivers relevant for digital preservation at SB, improve staff and management understanding of digital preservation challenges and enable SB to benchmark with other digital preservation organisations.

This paper presents the process and benefits of performing a self-assessment according to ISO 16363 at the State and University Library.

Firstly we introduce the content, organisation, policy framework and structure of SB's digital repository. Then we describe the assessment tools used in the last four to five years at SB. Finally the process of self-assessment according to ISO 16363 is described and evaluated.

2. DIGITAL PRESERVATION AT THE STATE AND UNIVERSITY LIBRARY

SB has been engaged in digital preservation for more than a decade. As a national legal deposit library responsible for collecting, preserving and disseminating audio-visual material it was quite early on clear to both head management and IT management that digital preservation would be a necessary investment and core objective for the library in the years to come.

SB is by law[1] obliged to collect and preserve broadcasted content from all Danish radio and TV channels and a representative cross section of all channels with production directed at a Danish audience. This legal deposit law for radio and TV was first passed in the Danish Parliament in 2005. From 2006 the library has collected this material using a combination of antenna and cable for digital preservation. In 2012 the library entered into an agreement with a service provider that delivers all the radio and TV material digitally to SB. Eight years of collecting radio and TV digitally means that the library now holds more than 2 PB of digital material, all stored in three copies, and growing with approximately 800 TB per year.

Adding to this is the Danish Net Archive established in 2005 in cooperation between The Royal Library, Denmark, and SB. Each institution holds a copy of the web archive. To date more than 400 TB have been collected.

Besides the radio and TV collection, the Newspaper Archive and the Net Archive, SB has a number of audio collections that have been digitized since 1999. These collections consist of rare and unique material often digitized from fragile media like wax cylinders and reel tapes. Also music and film material from ripped CDs and DVDs are preserved at SB. These collections range from small to medium sized (10 GB to 2 TB), but add to the complexity of the digital preservation task. Lastly the library collects digital cultural heritage material in the audio-visual area in general. All in all SB preserves very diverse collections, has very large amounts of data, and the repository is steadily growing in size and complexity.

2.1 Organising Digital Preservation

The National Library Division at SB is the formal owner of and thereby responsible for the preservation of all cultural heritage collections, including digital collections. In the early days of digital preservation at the library the preservation of digital collections were more or less considered the responsibility of the IT Division. A few years ago the National Library Division assumed the responsibility and is now in charge of decision-making for analogue as well as digital collections.

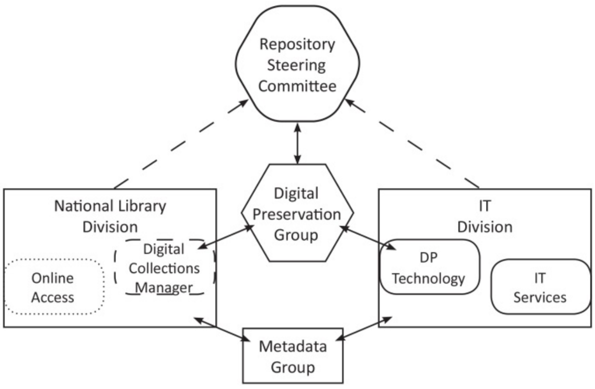

Repository Steering Committee
National Library Division
Digital Preservation Group
IT Division
Online Access
Digital Collections Manager
DP Technology
IT Services
Metadata Group

Figure 1 Organisation of Digital Preservation at SB

A policy [14] and a strategy [15] for digital preservation were created in 2011 and they clearly state who is responsible for what in connection with preserving the library's digital collections. At the same time a Digital Preservation Group (DP Group) with members from both the IT Division and The National Library Division was established. The DP Group discusses issues concerning digital preservation at the library and creates input for the library management concerning digital preservation matters, e.g. decisions on number of copies or choice of format for digitization projects etc.

In 2012 the National Library Division established a new function, the Digital Collections Management Team (DCM Team), with a Digital Collections Manager focussing on the digital collections in a number of different ways. The DCM Team collects information about all digital collections at the library, coordinates digital preservation actions with the IT Department and supports management's decision-making on digital preservation matters.

A Metadata Group for digital material was also established in 2012. This Metadata Group is concerned with metadata for digital collections and works with every digital collection that is digitized at the library, born digital or acquired externally. The Metadata Group is responsible for creating appropriate metadata schemas, collection metadata and supporting management when issues of buying, receiving or creating metadata are relevant. The Metadata Group also makes recommendations for the use of metadata with regards to online access to the digital collections.

Development of the technical infrastructure of the digital repository is primarily carried out in-house in the Digital Preservation Technology Department. The repository system is based on Fedora Commons (http://www.fedora-commons.org/) and the Bit Repository software (http://bitrepository.org/). The Digital Preservation Technology Department also creates tools and acts as consultants in the curation of the digital material.

2.2 Policy Framework

In order to perform digital preservation the best way possible SB has developed a policy framework for digital preservation and access to digitally preserved material. This framework consists of a number of documents that together supports the digital preservation work and decision-making in the library.

The policy framework includes the Digital Preservation Policy [14], Digital Preservation Strategy [15], Metadata

[1] Lov om pligtaflevering af offentliggjort materiale (Lov nr. 1439 af 22. december 2004), http://pligtaflevering.dk/loven/index.htm

Policy [16], Strategy for Information Channels [17] and the annual DS484 / ISO 27001-audit (Standards for information security) [3] and [7].

2.2.1 Digital Preservation Policy and Strategy

The library's Digital Preservation Policy [14] is a high-level policy supporting management's decision-making regarding digital preservation issues. A digital preservation strategy detailing the high level preservation policy into preservation procedure policies was developed in connection with the digital preservation policy. The Digital Preservation Policy and Digital Preservation Strategy [15] describe how digital preservation is to be carried out at SB.

The structures of the policy and the strategy documents are very similar. They consist of an Introduction and a Purpose section and a section defining the general framework for digital preservation including the library's aspirations for being a Trustworthy Digital Repository. Both documents also contain policy requirements on collection level. These policy requirements define issues such as how to manage bit preservation, which functional preservation strategies are preferred, how legal issues should be dealt with, what kind of QA is to be carried out etc. Especially this section is elaborated on in the Digital Preservation Strategy making it a very useful document in the literal sense of the word. The DP Group and the DCM Team use the Digital Preservation Strategy in daily collections handling and decision-making. It is a key point that the Digital Preservation Strategy is not just an act of intention but is in fact acted on.

On the basis of the Digital Preservation Strategy the digital collections management team has created *collection plans* for each digital collection. These plans reflect policy requirements in the strategy enabling the team to add information about the collection and decisions made for the collection. The collection plans are created and stored in a wiki accessible by all SB library staff. The plans are updated whenever new decisions or materials are added and are reviewed once a year.

SB is a partner in the SCAPE project (www.scape-project.eu) and has as such been deeply involved in the policy guidelines work in SCAPE, [12] and [13] based on our experience with policy work at the library.

2.2.2 Metadata Policy

Being a national library metadata is an issue and SB has created a general Policy for Metadata [16] including metadata for digital material. The Metadata Group for digital material is carrying out their work concerning metadata for digital material on the basis of this policy.

2.2.3 Strategy for Information Channels

Digital preservation should also take the question of access to the collections from the designated community or communities into account. As part of the policy framework for digital material a Strategy for Information Access [17] has been developed in the library. This policy describes in large how and which channels the library will use in providing access to the digital material for its designated communities.

2.2.4 Standards for Information Security

As every national institution in Denmark, SB has until 2013 been obliged to audit the organisation using DS484 [3], a Danish standard for information security, shifting to ISO 27001 [7] from 2014 onwards. The annual audit of information security at the library constitutes the basis for the policies and strategies concerning digital material at the library and includes an inventory of information assets.

2.3 Digital Preservation Infrastructure

SB supports open source software and the infrastructure for digital preservation is built upon open source software components. The digital infrastructure including the repository is basically comprised of two closely linked systems: one for bit preservation and one for functional preservation.

2.3.1 Bit Repository

The Bit Repository at SB has been developed in cooperation with the Danish State Archives and the Royal Library of Denmark. It is described as "The purpose of the Bitrepository system is to enable longterm preservation of data in a distributed, highly redundant architecture. The data integrity is ensured by using multiple, independently developed data storage systems (…)" (www.bitrepository.org).

SB has two geographically independent locations for data storage and at the same time ensures that organisational responsibility is divided between independent units/persons at the library in order to secure the bit preservation.

SB operates with a number of different levels for bit preservation. In order to be able to stringently determine the necessary bit preservation level for each collection, a bit preservation level scheme has been created. This scheme is used for assessing all new collections regarding number of copies, geographical location of copies and level of bit integrity checking. The assessment is performed by examining the collection and judging its value whereby determining which bit preservation level should be used. For example, the digital preservation of the radio and TV collection is performed according to national legal deposit law and the obligation to keep it safe for the future. This digital collection has no physical counterpart and is thereby a unique national collection. It was therefore decided that this collection will be kept in three copies placed in three different data storage systems at the two locations provided by the library and is to the greatest extent possible preserved by using different technologies. But the size of the collection sets a limit for the bit preservation effort and it has been decided not to have an online preservation copy due to huge economic expenses. Therefore the collection is preserved in two offline tape copies and 1 nearline tape copy but as far as possible on tapes from different providers to avoid erroneous tape batches.

2.3.2 Metadata Repository

SB preserves metadata and performs functional preservation using an in-house developed Digital Object Management System (DOMS). This system is built on the open source Fedora Commons system. The library supports the continuous development of Fedora Commons by having a developer assigned as part of the Fedora Commons development team with commit privileges. Metadata for digital collections are preserved in this system with linkage to the files in the bit repository. In the document Digital Preservation Strategy [15] it is stated that the preferred functional preservation strategy is to migrate only when files in a given format are endangered. So far SB has not needed to perform migration for any collections in the repository.

2.4 www.digitalbevaring.dk – a Forum for Digital Preservation in Denmark

Over the last few years SB has strived to professionalise the field of digital preservation in the library. In that process the library has obtained a lot of useful experience that could benefit other organisations concerned with digital preservation in Denmark. Therefore SB in cooperation with the Danish State Archives and the Royal Library of Denmark has established the website www.digitalbevaring.dk (in Danish). The website consists primarily of articles about digital preservation and digitization issues that the three institutions have had experiences with or obtained knowledge of. Cooperation about the website content has proved very useful in the attempt to establish common definitions in Danish of digital preservation issues amongst the large, national cultural heritage institutions in Denmark.

3. THE ROAD TO TRUSTWORTHINESS

3.1 A Brief History of Trustworthiness

The initial work on trustworthiness started with the late 1990's 'OAIS-compliancy' [2] and the work of the RLG and OCLC working group, which published Trusted Digital Repositories: Attributes and Responsibilities [10], a document which has provided helpful recommendations and guidance to institutions struggling with digital long-term preservation. The growing interest in organising the work on digital preservation led to a task force on trustworthiness[2] in digital repositories. This task force published TRAC [1] which defines a Trusted Digital Repository as one whose 'mission [is] to provide reliable, long-term access to managed digital resources to its designated community, now and into the future'[3]. Other initiatives on trustworthiness were undertaken through the 2000's in Europe [4], [5] and [11].

A survey conducted in the CASPAR project [6] concluded that "evidence of previous effective curation and conformity to international standards are the most important factors in determining whether to trust a repository". These conclusions underpin the importance of adhering to international standards and justify a standard on trustworthiness.

In 2012 an ISO standard, ISO 16363 - Space data and information transfer systems – Audit and certification of trustworthy digital repositories [8] was published. This standard is based on TRAC and defines a recommended practice for assessing the trustworthiness of digital repositories.

3.2 The Concept of Trustworthiness at SB

SB's work on trustworthiness of digital repositories began with the EU FP6 project DigitalPreservationEurope (DPE, http://www.digitalpreservationeurope.eu) in which SB was co-author of the Planning Tool for Trusted Electronic Repositories (PLATTER) [11] and the DRAMBORA toolkit for self-assessment [4].

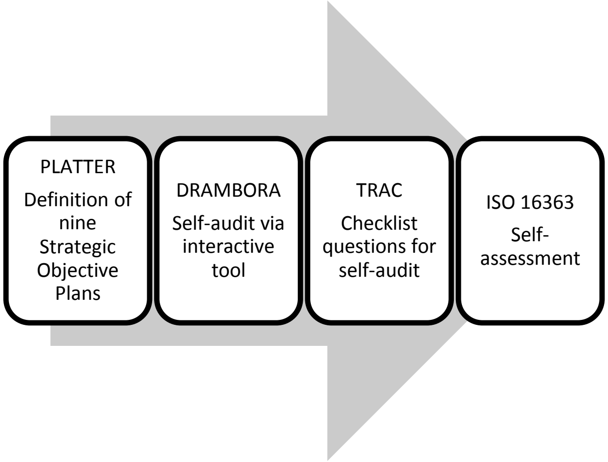

PLATTER
Definition of nine Strategic Objective Plans
DRAMBORA
Self-audit via interactive tool
TRAC
Checklist questions for self-audit
ISO 16363
Self-assessment

Figure 2 Illustration of SB's road to Trustworthiness

As a natural consequence of this engagement PLATTER and DRAMBORA were chosen as means for the first self-assessment work at SB in 2010/11. Led by a small group of organisational specialists the PLATTER toolkit was inspected and a guideline was written for each of the nine suggested PLATTER plans. This work involved both technical and organisational specialists. Based on the nine plans the DRAMBORA toolkit was used to define objectives, mandates and constraints. A total of 78 risks were identified within technical, administrative and organisational fields. All risks were then assessed and the importance of the most severe risks was stressed in an internal report to the SB Management.

Concrete results of this work (2011) were decisions to

- appoint a Digital Collections Manager
- create and maintain an annual business plan for the repository
- intensify the focus on knowledge sharing and documentation

Besides the decisions mentioned above a list of tasks to mitigate the risks identified in the DRAMBORA exercise was developed. This list formed the "stepping stones" for the ensuing work on developing the field of digital preservation at the library and for the work carried out in the DP Group.

3.3 Competency Development

In 2012 SB were offered to participate in a week-long course on digital curation organised by the American initiative DigCCurr from the University of North Carolina at Chapel Hill (http://www.ils.unc.edu/digccurr/aboutII.html#dce). This course offered an introduction to some of the main areas of digital curation, including work and tools for working with trustworthiness. Both authors of this paper participated in the course which inspired us to perform a new self-assessment of SB's digital repository but this time based on the ISO Standard 16363 [8].

Additionally, one of the authors participated in 'Trust and Digital Preservation' (www.dpconline.org/events/details/61-trust-and-digital-preservation), a two-days training event in

[2] Joint task force to address digital repository certification - Research Library Group (RLG) and the National Archives and Records Administration (NARA)

[3] [1] TRAC p. 3

Dublin, 2013, at which an overview of audit and certification was given and each participant attempted to fill out the Data Seal of Approval (http://www.datasealofapproval.org/en/).

3.4 Deciding on an Audit Strategy

In SB's Digital Preservation Strategy it is stated that SB "*seeks to achieve the status of Trustworthy Digital Repository and thereby meet internationally acknowledged standards*" [15]. The library will perform an audit every second year and it will be decided from audit to audit whether it should be an internal process or an external audit with certified auditors visiting the library. In 2010/11 the library chose the tools DRAMBORA [4] and PLATTER [11] for a first self-assessment.

Working with DRAMBORA and PLATTER broadened the audit team's knowledge of the structure of SB's repository and how the different tasks related to digital preservation are organised at SB. The 'DRAMBORA interactive' was easy to use, and filling out the preparation material provided an overview of the different functions, responsibilities and roles at the library. By examining the PLATTER plans written and by questioning staff we were able to identify 78 risks. Once the risks were all assessed a list of all risks rated by Risk Probability, Risk Impact, and Risk Severity was generated. However, this Risk Register could only be extracted as a pdf, and it was very difficult to relate risks to other risks. We received an extract of the database, but again, this resulted in lists that could not easily be related to each other thus obstructing easy handling of the input to DRAMBORA for further use

In 2012/13 it was decided to base the coming audit on the new ISO 16363 for Certification of Trustworthy Digital Libraries [8] and perform a self-assessment. This decision was made based on an assessment of the certification readiness of the organisation. As soon as the ISO standard was published the DP Group realised that doing an external audit would be a future process due to the amount of work and the organisational maturity.

3.5 Self-assessment using ISO 16363

On the website http://www.iso16363.org/ a process for preparing for an external audit using ISO 16363 is described. This process contains steps such as answering all metrics, produce evidence for all metrics etc. It was decided to do a self-assessment focused on establishing a process for auditing, answering metrics, and produce a substantial reference list.

The purpose of performing a self-assessment of the digital preservation of the library using ISO 16363 was to work thoroughly through all processes, workflows, systems and organisational build up to be able to expose all drivers relevant for digital preservation at SB. At the same time a self-assessment would be part of improving staff and management understanding of digital preservation challenges and form a basis for future competency development planning in digital preservation. A self-assessment would with the intensive internal review and the substantial reference/evidence list produced provide a steady and well-known ground to further develop digital preservation activities at the library. Finally a self-assessment (and when time comes an external audit) would enable SB to benchmark with other digital preservation organisations around the world and enhance cooperation and common development in the area of digital preservation.

4. Carrying out the Self-assessment

ISO 16363 [8] is split into three main sections which provide the normative metrics against which a digital repository may be evaluated. The sections are:

- Organisational Infrastructure
- Digital Object Management
- Infrastructure and Security Risk Management.

Each section has a number of metrics, like 'The repository shall be able to identify which definition applies to which AIP' (#4.2.1.1) with supporting text, examples and discussion to facilitate the process.

Initially a tailor-made wiki was agreed to be an appropriate tool for documenting the conclusions to the metrics. We started the project by making a wiki page for each of the 109 metrics with the ISO standard's texts, both Supporting text, Examples and Discussions. The idea was to add all information regarding a specific metric to its wiki page and then aggregate excerpts on special pages, but it was difficult to keep a sense of perspective in the daily work with metrics scattered on a large number of pages.

Thus, it was decided to use the PTAB[4] spreadsheet [9] which is divided into three pages, each representing the metrics of one of the main sections and grouped into one or more subsections. We added extra columns on each sheet to be able to add comments and also ratings based on the rating system from the Drupal TRAC review tool by MIT[5].

4.1 Understanding the Metrics

The DCM Team at SB started out by reading the ISO 16363 thoroughly and discussing each metric. This formed the basis for selecting library staff with knowledge of the infrastructure of SB's digital repository, including both the metadata repository (DOMS), the data repository (bit repository), and the overall organisational aspects of the library.

A group of four people, the DCM Team together with two IT developers, then worked their way through all the metrics. Each metric was discussed by the group and an explanation of how SB fulfils the metric was added. This work did not have a dedicated time period or time frame assigned to it. It was performed in and between meetings up to three weeks apart. This means that the assessment period has been quite long, and some metrics were so abstruse in our understanding, that the group sometimes found it difficult to recognise an explanation to a specific metric the next time we met. So this long stretched process has occasionally made reiterations necessary.

Understanding and agreeing on the actual meaning of the metrics proved to be a difficult task, due to the fact that some metrics were difficult to adapt to the organisation at SB. Additionally the language barrier turned out to be more difficult to overcome than expected.

[4] The Primary Trustworthy Digital Repository Authorisation Body (PTAB)

[5] https://www.archivematica.org/wiki/Internal_audit_tool#Drupal_TRAC_review_tool

Several metrics were of a kind that the four members of the group were not in a position to answer themselves, so many other people have been involved; specialists, managers and for more clarifying questions also one of the initiators of the Drupal TRAC review tool, Nancy McGovern (Curation and Preservation Services at MIT Libraries).

Once all metrics were described (see example in Table 1), the Digital Preservation Architectural Team at SB reviewed the explanations of how the metrics were met. They presented their comments to the group for discussion and this lead to minor revisions to better describe the processes within and infrastructure of the repository.

Table 1 Metrics example

Metric 3.3.6	THE REPOSITORY SHALL COMMIT TO A REGULAR SCHEDULE OF SELF-ASSESSMENT AND EXTERNAL CERTIFICATION.
Explanation of how the repository addresses this metric	The schedule for self-assessment is stated in REF004 DP Strategy and in REF044 DCM Annual Cycle. Results of self-audit in 2010 can be seen in REF020 DRAMBORA Report. Results from 2012-13 can be seen in REF072 TDR wiki (internal).
Brief description of evidence	REF004 DP Strategy REF018 Platter REF020 DRAMBORA Report REF021 Audit Planning REF044 DCM Annual Cycle REF072 TDR wiki

4.2 Reference List

As evidence for each metric a list of titles of existing documents that describes policies, procedures, and practices at the SB relevant to the metric was made concurrently with the self-assessment process (see example in Table 1, lowermost row). For each explanation to a metric the relevant documents were recorded with ID numbers and short title. This list serves as evidence that the repository is complying with the metric described. A more detailed description of the documents is provided in a separate sheet (the Reference tab) of the PTAB spreadsheet [9].

This 'Reference list' is now a very comprehensive and helpful tool for the digital preservation work at the library as it includes all documents mentioned earlier in the policy framework chapter as well as descriptions of procedures, workflows, processes, software documentation etc. In the list a link for each document is provided as well as the name of the person responsible for maintaining the document.

During the process several areas were identified that need further documentation and these have also been listed in the Reference list but marked as 'not written yet'. Together with other tasks identified during the self-assessment process these are now listed in a task list with assigned task managers.

The uncovering of evidence has been an extremely valuable process and has resulted in a number of concrete tasks. It is now very clear in what parts of the digital repository additional documentation and workflow descriptions etc. are needed.

4.3 Responsibilities

A list of staff involved in digital preservation has also been compiled during the self-assessment work. For each metric the staff member responsible for the explanation of compliancy to the metric has been identified and for each reference a staff member is identified as being responsible for keeping the specified document up to date. This leaves SB with a clear understanding of joint and divided responsibilities in digital preservation at the library.

4.4 Compliance Rate

The self-assessment process at SB also included rating each compliancy explanation according to the compliance rating system from the Drupal TRAC review tool[6].

The Drupal TRAC review tool defines five levels of compliance:

- 0 = non-compliant
- 1 = slightly compliant
- 2 = half compliant
- 3 = mostly compliant
- 4 = fully compliant

Compliance rates provide an easy overview of the state of the repository. We decided not only to define a compliance rate but also a 'compliance wish' showing how high a rating we would like SB's repository to achieve for the specific metric. The additional compliance wish reveals how compliant we think SB can – and desires to – become within its budgets, organisational framework and digital preservation goals. SB does not always wish for a rating as 'fully compliant' due to the fact that SB as a digital repository does not match the ISO standard one-to-one.

A low compliance rate would indicate that a metric is not fulfilled, but if the compliance wish is also low, SB has no intention of increasing the rating in the near future. If, on the other hand, the compliance rate is low, but the compliance wish is high, this indicates an area which needs special attention. An explanation for the compliance wish has been inserted in the spreadsheet whenever the compliance wish diverts from 'fully compliant'.

To shorten the discussions of rating the metrics we used a set of 'Planning Poker' cards known from SCRUM sprint planning[7]. A plain discussion of each metric would 1) have taken a long time and 2) easily result in people changing their immediate choice when they hear how the others rate a specific metric. To avoid this each member of the group was given a set of cards ranging between 0 and 4 and then 'played' the number they found most appropriate in terms of how compliant the member thought that SB is compared with the ISO 16363 requirements. If a metric rating returned four identical cards there was no need for further discussions. Whenever the cards 'played' were not identical a quick

[6] https://www.archivematica.org/wiki/Internal_audit_tool#Drupal_TRAC_review_tool

[7] http://en.wikipedia.org/wiki/Planning_poker

discussion led to a common understanding or if necessary an elaboration of the metric. The 'rating' poker exercise left us with a clear and common understanding of how SB meets the metrics of ISO 16363 and concluded the self-assessment very effectively.

4.5 Challenges with ISO 16363

Some of the obstacles or challenges we met while working with ISO 16363 are described below.

We experienced that section 3 Organisational Infrastructure required a lot of documents and plans to be presented. Most of this material is already part of the SB digital repository set up and thus easy to answer and provide evidence for. Section 3 also contains questions about monitoring which we found more difficult to answer as it did not seem all that clear what evidence would be sufficient. Other examples of where it seems difficult to produce a sufficient answer would be questions concerning "Staff with adequate skills and experience" - how do you determine if SB has hired the right staff? We believe that we have the right staff and we have staff exam papers etc. to prove it but is that sufficient? The same goes for a metric asking for evidence that the organisation has "the appropriate number of staff". Appropriate number of staff is difficult to answer unequivocally and depends highly on who you ask - the financial department or the head of IT.

Section 4 Digital Object Management was especially challenging when it came to the language barrier. Long discussions took place about the precise definition and translation of terms. Also discussions about intangible terms such as "appropriately verify" and "sufficient control" took place. What does it take to "verify appropriately" or "control sufficiently" when you strive to be a trustworthy digital repository?

In section 5 Infrastructure and Security Risks the discussions and challenges evolved primarily around the level of detail. A lot of the metrics that were debated in this section were about systems and procedures and systems to monitor systems. It took time to define the level of detail for each metric. We have systems that monitor our systems and procedures for acting on notifications etc. But should it be as detailed as describing the procedures for ensuring that the procedures are followed?

There were quite a lot of supporting text to be found in the further descriptions of the metrics and in the examples of evidence but as described above a great deal of the metrics still posed challenges.

In general the challenges we met led to fruitful discussions but were also very time consuming.

5. LESSONS LEARNED

Working with ISO 16363 turned out to be a challenge for the library. The experience from earlier work with DRAMBORA etc. was helpful but using the ISO 16363 was a very different way to work with self-assessment. Being forced to making implicit knowledge, processes and workflows explicit and prove the trustworthiness of the digital repository by producing evidence for all statements has been a complicated and laborious task. All in all the amount of work put into the self-assessment over a period of 15 months sums up to three full person-months. The DCM Team has used the main part of these hours but also technical staff has been involved in varying degree. SB had not dedicated a specific time period for the self-assessment or stated a deadline for the work which means that the self-assessment stretched over a long period of time. It would probably have been more efficient to dedicate a shorter but more intense period of time working on the self-assessment. This aspect will of course be considered when we start planning for the next round of assessment. The benefits of the self-assessment have been numerous – our understanding of all aspects of the digital repository has grown substantially, and the correlations and interdependencies between the different parties and tasks at the library have become much more transparent.

6. NEXT STEP

The self-assessment has been summarised in an internal report with conclusions of the work. In this report the most important recommendations for the management to consider here and now will be specified together with the long term considerations.

The self-assessment has enlarged our insight in many of the more specific procedures within the digital preservation of SB's digital repository. A list of things-to-do has been created in parallel with the self-assessment audit and this includes both improved documentation of specific processes, policies that need to be written or edited, and preservation procedures that are not carried out the best way possible as it is now. Where possible a task manager has already been assigned to each task during the self-assessment process, and deadlines for the tasks will be added together with additional task managers in the near future.

7. CONCLUSION

After the very comprehensive work of performing a self-assessment of SB we are left with a valuable snapshot of how SB is performing as a repository for digital material. Both organisational and technical solutions have been thoroughly examined and we have obtained a common view on SB as a digital repository. The self-assessment has led to a fuller understanding of a common vision for digital preservation between the different parts of the organisation. At the same time the self-assessment has worked as competency development for the staff involved in the exercise. The self-assessment has produced a general organisational awareness concerning digital preservation and the demands for trustworthiness. It has also produced a gap analysis and has helped identify a number of tasks that will be used for further development of SB as a trustworthy digital repository.

The self-assessment has led to a number of specific tasks to be carried out to improve SB's digital repository. As SB has had its focus on optimising digital preservation procedures – technically and organisationally – for almost a decade this self-assessment has not led to any substantial changes in the organisation as such but acts as a new baseline to build future improvements on and is as such a very valuable tool for the organisation.

A major task after concluding the self-assessment is to transfer the knowledge and results produced during the self-assessment to daily work enhancing digital preservation at the library. This will be done by clearly communicating the results of the self-assessment, including clarifying who is responsible for updating evidence and performing the tasks identified. The DP Group and the DCM Team are in charge of following up on tasks and evidence.

SB's Digital Preservation Strategy [15] states that an audit must be performed every second year to keep the organisation fit. This time the audit was performed as a self-assessment which has proved to be a valuable and comprehensive method of evaluating SB as digital repository. The aim is, in time, to be certified as a trustworthy digital repository but the process with conducting the self-assessment using ISO 16363 has revealed that there is still work to be done at the library before we are ready for external auditing.

SB chose to perform a self-assessment and is thus not a certified trustworthy digital repository. We do consider ourselves trustworthy, though, in the sense that we made a self-assessment that identifies every step, action and piece of evidence in the long term digital preservation at SB, both organisational and technical. As part of being trustworthy SB publishes non-confidential material online, e.g. the policy framework is available from http://www.statsbiblioteket.dk, and software documentation is available from code repositories such as GitHub (https://github.com/). The confidential material concerning digital preservation at SB is available for those whom it may concern.

8. ACKNOWLEDGEMENTS

The authors would like to thank Head of Digital Preservation Technology, Bjarne S. Andersen and IT Project Manager Lars F. Mydtskov for their participation in and input on the underlying study. Their comments, feedback and evaluation helped clarify technical issues and conclude the self-assessment. Any remaining misinterpretations or mistakes are those of the authors.

9. REFERENCES

[1] Center for Research Libraries. *Trustworthy Repositories Audit & Certification: Criteria and Checklist.* Version 1.0. February 2007. Chicago: CRL www.crl.edu/PDF/trac.pdf

[2] Consultative Committee for Space Data Systems. *Reference Model for an Open Archival Information System (OAIS)*, Magenta Book, 2012, CCSDS Secretariat

[3] Danish Agency for Digitisation. *Standard for information security*. Retrieved 21-03-2014 from http://www.digst.dk/Servicemenu/English/IT-Security/Standard-for-information-security.aspx

[4] Digital Curation Centre (DCC) and DigitalPreservationEurope (DPE), *DRAMBORA*. 2008. http://www.repositoryaudit.eu/

[5] Dobratz, S.; Schoger, A. and Strathmann, S. *The nestor Catalogue of Criteria for Trusted Digital Repository Evaluation and Certification*, Journal of Digital Information 8, no. 2 (2007) http://journals.tdl.org/jodi/index.php/jodi/article/view/199/180

[6] Giaretta, D. et al. *Report on Trusted Digital Repositories*. 2009. CASPAR Consortium. http://www.alliancepermanentaccess.org/filestore/CASPAR-deliverables/CASPAR-1203-RP-0101-1_0.pdf

[7] ISO. *Information Security Management (ISO/IEC 27001:2013).* International Standards Organization, 2013 http://www.iso.org/iso/home/standards/management-standards/iso27001.htm

[8] ISO. *Space data and information transfer systems – Audit and certification of trustworthy digital repositories (ISO 16363:2012(E)*. International Standards Organization, 2012 http://www.iso16363.org/

[9] PTAB. *Self-Assessment Template for ISO 16363,* http://www.iso16363.org/assets/Self-AssessmentTemplateforISO16363.xls

[10] Research Libraries Group. *Trusted Digital Repositories: Attributes and Responsibilities*. An RLG-OCLC Report. Mountain View, CA: RLG, May 2002 http://www.oclc.org/content/dam/research/activities/trustedrep/repositories.pdf

[11] Rosenthal, C. et al. *D3.2 Repository Planning Checklist and Guidance, DigitalPreservationEurope (DPE),* http://www.digitalpreservationeurope.eu/platter/

[12] Sierman, B., Jones, C., Bechhofer, S. and Elstrøm, G. 2013. Preservation Policy Levels in SCAPE. Reasoning about naming systems. In *Proceedings of the 10th International Conference on Preservation of Digital Objects* (Lisboa, PT, 2013), Biblioteca Nacional de Portugal, 225-230

[13] Sierman, B., Jones, C. and Elstrøm, G. *Catalogue of preservation policy elements.* D13.2 Draft. 2014 http://www.scape-project.eu/wp-content/uploads/2014/03/SCAPE_D13.2_KB_V1.0.pdf

[14] Statsbiblioteket. *Digital Preservation Policy for the State and University Library Denmark. Version 2.0.May 2012.* Retrieved 21-03-2014 from https://en.statsbiblioteket.dk/about-the-library/ddpolicy

[15] Statsbiblioteket. *Digital Preservation Strategy for the State and University Library Denmark. Version 2.0 June 2012.* Retrieved 21-03-2014 from https://en.statsbiblioteket.dk/about-the-library/dpstrategi

[16] Statsbiblioteket. *Metadatapolitik for Statsbiblioteket. Version 25. februar 2013.* Retrieved 21-03-2014 from http://www.statsbiblioteket.dk/om-statsbiblioteket/publikationer/statsbibliotekets-metadatapolitik

[17] Statsbiblioteket. *Strategi for Informationsforsyning*. Version 1.2 14. maj 2012. Retrieved 21-03-2014 from http://www.statsbiblioteket.dk/om-statsbiblioteket/filer/stragegi_for_informationsforsyning

Decommissioning of legacy systems: A methodology for indentifying and preserving records of ongoing business value in legacy business systems

Ingrid MacDonald
61 7 3131 7729
ingrid.macdonald@archives.qld.gov.au

Adrian Cunningham
61 7 3131 7957
adrian.cunningham@archives.qld.gov.au

Anna Morris
61 7 3131 7767
anna.morris@archives.qld.gov.au

Neal Fitzgerald
61 7 3131 7982
neal.fitzgerald@archives.qld.gov.au

Queensland State Archives
435 Compton Rd, Runcorn
Queensland, AUSTRALIA

ABSTRACT

This paper provides an overview of the methodology and supporting advice developed by Queensland State Archives to help Queensland government departments undertaking the disposal or replacement of legacy ICT assets to lawfully manage the information content of those systems.

Before deleting or investing in any preservation effort on records in decommissioning candidate systems, agencies need to have a clear understanding of:

- which records need to be kept and for how long
- how to seek disposition authorisation for records which may have no ongoing value to the organisation, and
- what is and is not deemed a suitable preservation environment for managing records of ongoing value, from a recordkeeping perspective.

Developed with the non-record professional in mind, a key message underpinning the methodology is that records as evidence of business activity are strategic assets in their own right, not just a byproduct of the business process, owned by the application. [Gartner 2011]

General Terms

Strategic environment, preservation strategies and workflows, case studies and best practice.

Keywords

Legacy business systems, Methodology, Disposition, Preservation

1. INTRODUCTION

In 2012, the Queensland government undertook an audit of significant ICT assets across the 20 Queensland state government departments. The Audit, the first of its kind in Queensland, was undertaken by the Queensland Government Chief Information Office and identified that a large number of central government business systems are run on unsupported, or soon to be unsupported technology: presenting a high risk to government.

2012 ICT Audit - Headline findings [Queensland Government. 2012]

- 1730 systems reported to the ICT systems audit
- 904 (54%) are "legacy ICT systems", potential candidates for rationalisation or decommissioning
- 91% of the 1730 systems will be at end of life within 5 years

In response to the audit findings, the Queensland government set out to realise cost savings of up to AUD 10 million per annum by decommissioning legacy ICT systems no longer in active use. The Audit and its recommendations largely focused on the systems: remaining largely silent on the value of their information content.

Queensland State Archives role as lead agency for recordkeeping signaled to Government that most of these legacy business systems contain public records which must be managed in accordance with the *Public Records Act 2002* (the Act) and set out to answer two questions for impacted agencies: Which records could be legally disposed of and how? and Which records need to be retained and preserved?

The ICT Audit and the subsequent drive to rationalise a substantial number of legacy business systems has brought to the forefront digital records preservation and disposal issues. Issues which systems administrators, Chief Information Officers and others charged with the commissioning and decommissioning of ICT assets within an organisation typically do not engage with.

2. EXISTING METHODOLOGIES AND FRAMEWORKS

Before developing new tools, Queensland State Archives examined the suitability of any existing methodologies and frameworks which could be referenced either from within the Queensland government, or elsewhere.

The Queensland government's *Application Rationalization Methodology* (ARM) was developed by the Queensland Government Chief Information Office. Used primarily by IT personnel and focusing on systems, the methodology does include a chapter on recordkeeping obligations which had been developed in consultation with Queensland State Archives. At the time of its original publication however, key aspects of Queensland State Archives appraisal and sentencing (selection) of records in business systems were under review. Consequently the ARM provides little practical guidance on the application of disposition authorities to digital records.

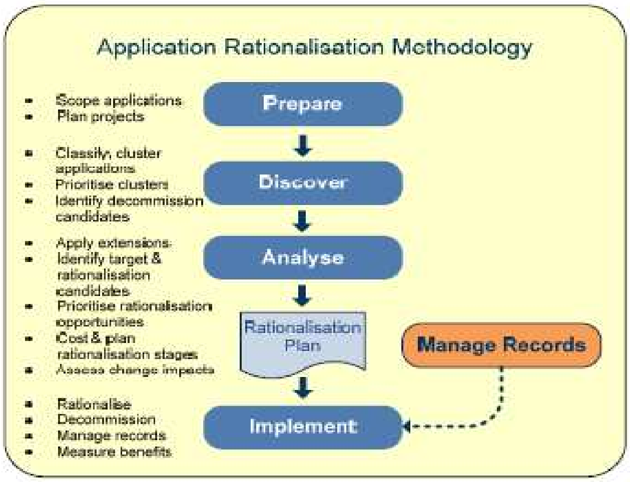

Application Rationalisation Methodology
Scope applications
Plan projects
Prepare
Classify, cluster applications
Prioritise clusters
Identify decommission candidates
Discover
Apply extensions
Identify target & rationalisation candidates
Prioritise rationalisation opportunities
Cost & plan rationalisation stages
Assess change impacts
Analyse
Rationalisation Plan
Manage Records
Rationalise
Decommission
Manage records
Measure benefits
Implement

Figure 1: Queensland Government Chief Information Office Application Rationalisation Methodology [Queensland Government Chief Information Office. 2011]

While appraisal and disposition authorisation processes at Queensland State Archives were well established before the 2012 Audit, the potential for disposition authorisation requests arising from agencies potentially under pressure to decommission systems presented two particular challenges for Queensland State Archives:

- Queensland State Archives' capacity to process a large number of requests for disposition authorisation quickly
- The potential expectation which agencies might have that authorisation would be given, given the priority given to the issue by government.

For these reasons, Queensland State Archives recognised that a number of artifacts were required to meet this demand:

- A methodology that clearly set out the expectations and obligations which all agencies were expected to meet
- A transparent and defensible set of criteria which would be used to assess any applications for disposal and that would stand up to public scrutiny
- Mechanisms for seeking special consideration for certain types of records and for reporting records that were already lost
- Advice to supplement known gaps in Queensland State Archives existing policy framework which brought together relevant advice 'under one roof'.

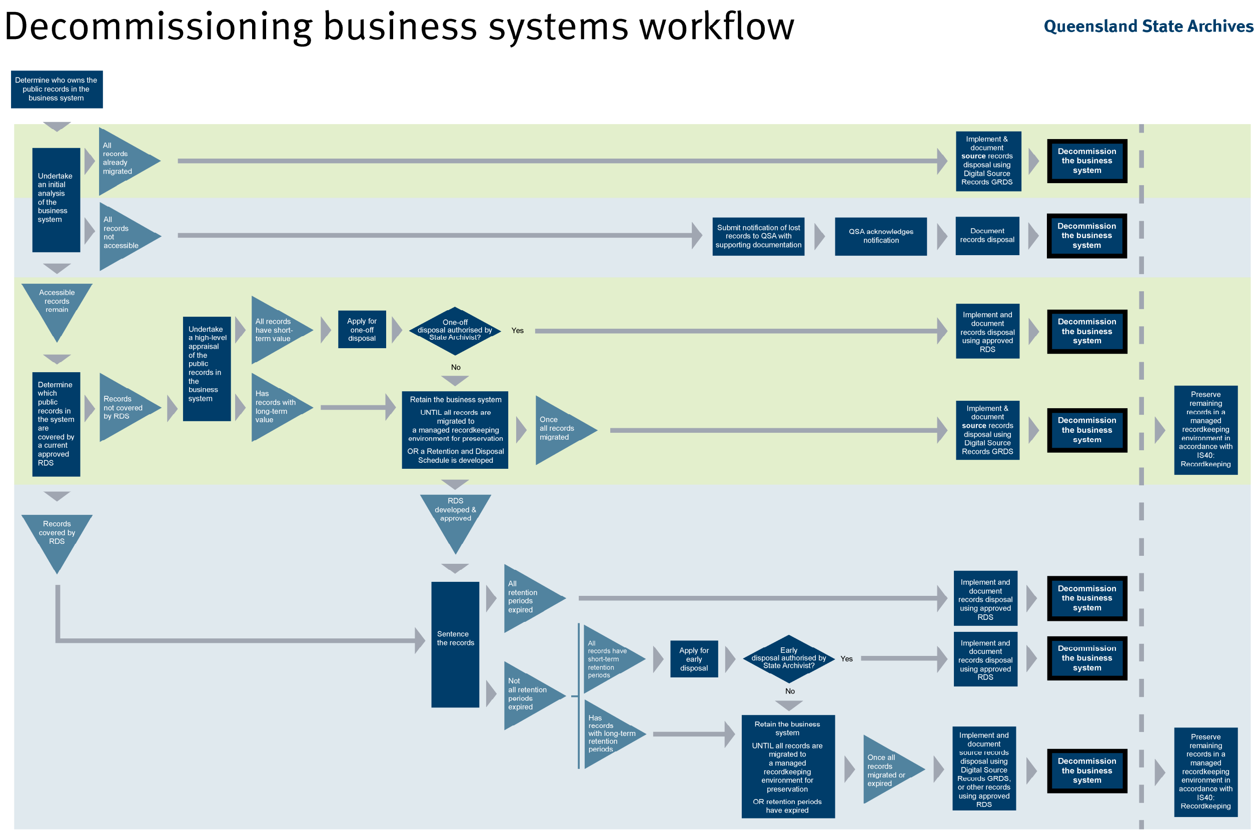

Decommissioning business systems workflow
Queensland State Archives
Determine who owns the public records in the business system
Undertake an initial analysis of the business system
All records already migrated
Implement & document source records disposal using Digital Source Records GRDS
Decommission the business system
All records not accessible
Submit notification of lost records to QSA with supporting documentation
QSA acknowledges notification
Document records disposal
Decommission the business system
Accessible records remain
Undertake a high-level appraisal of the public records in the business system
All records have short-term value
Apply for one-off disposal
One-off disposal authorised by State Archivist?
Yes
No
Implement and document records disposal using approved RDS
Decommission the business system
Determine which public records in the system are covered by a current approved RDS
Records not covered by RDS
Has records with long-term value
Retain the business system
UNTIL all records are migrated to a managed recordkeeping environment for preservation
OR a Retention and Disposal Schedule is developed
Once all records migrated
Implement & document source records disposal using Digital Source Records GRDS
Decommission the business system
Preserve remaining records in a managed recordkeeping environment in accordance with IS40: Recordkeeping
RDS developed & approved
Records covered by RDS
Sentence the records
All retention periods expired
Implement and document records disposal using approved RDS
Decommission the business system
Not all retention periods expired
All records have short-term retention periods
Apply for early disposal
Early disposal authorised by State Archivist?
Yes
No
Implement and document records disposal using approved RDS
Decommission the business system
Has records with long-term retention periods
Retain the business system
UNTIL all records are migrated to a managed recordkeeping environment for preservation
OR retention periods have expired
Once all records migrated or expired
Implement and document source records disposal using Digital Source Records GRDS, or other records using approved RDS
Decommission the business system
Preserve remaining records in a managed recordkeeping environment in accordance with IS40: Recordkeeping

Figure 2: Queensland State Archives' Managing public records when decommissioning systems workflow

3. ABOUT THE METHODOLOGY

Queensland State Archives' Decommissioning Methodology [Queensland State Archives 2013a] comprises an interactive workflow, supported by a suite of advice around the core challenges of:

- identifying if a system contains public records
- managing the separate requirements of public records in the same system requiring temporary or permanent retention
- managing the disposal process for public records which have not yet reached their minimum retention period or which are not yet covered by a Retention and Disposal Schedule approved by the State Archivist
- determining the most appropriate digital preservation strategy for the public records.

Depending on a number of variable factors, for some agencies these challenges will be easily or already resolved, for others the agency would need to undertake more detailed analysis of the system and its content. These variables are depicted in figure 2 above as four different disposition/preservation scenarios or pathways. A higher resolution version of the document can be found on the Queensland State Archives web site[1].

In the first scenario, all records have already been migrated to another system - that is, the system contains 'copies' of data no longer relied on as the record of the agency but which may have been retained because the agency lacked confidence that the copies could be lawfully deleted. Agencies in this scenario are able to delete the source records without further approval from the State Archivist provided they have met all migration conditions in Queensland State Archives' *General Retention and Disposal Schedule for Digital Source Records* [Queensland State Archives 2012].

The second scenario deals with the situation where all records in the business system are no longer accessible, that is, they can no longer be opened or interpreted. Conscious that this pathway could be used as a potential easy option by agencies unwilling to invest in the ongoing management of the records, Queensland State Archives nevertheless acknowledged that there may be a limited number of legitimate cases where the records in the systems were already effectively lost due to some catastrophic system failure or obsolescence. To guard against this, a number of checks and balances were built into the disposal approval process. For example, agencies are required to notify Queensland State Archives of the circumstances surrounding the 'loss' prior to their deletion. Agencies in this scenario cannot delete the records without first providing evidence to the State Archivist that the records are irretrievable.

The next scenario deals with records in a business system that are still accessible and are covered by a current disposition authority approved by the State Archivist which sets out the minimum legal retention period for the records in the system. Agencies in this scenario are able to delete the records in the system without seeking further authorisation from the State Archivist if their minimum authorised retention periods have expired. However their deletion must be approved by the legal owner of the records – that is, the agency which owns the function to which the records relate, not the IT system owner. For those records which have not met their minimum retention period, agencies need to consider the best way to preserve the records for the remaining retention period. In some cases, permanently.

The final scenario deals with those records in systems which are not covered by a current disposition authority and therefore the minimum legal retention period for the records is unknown. Because under Queensland's *Public Records Act 2002* records cannot be disposed of without authorisation, agencies in this scenario have the option to either undertake an appraisal and disposition authorisation exercise, or make arrangements for the preservation and management of the records.

Of course, these scenarios are not necessarily mutually exclusive. For example, it is possible that a system might contain records, some of which are covered by an existing disposition authority and others that are not. In such cases, more than one pathway of the methodology may need to be followed to finally determine if the records are for disposal or preservation candidates. To keep the workflow diagram as simple as possible, all possible variations have not been depicted.

3.1 Key Issues Discussed in the Toolkit

3.1.1 Do business systems contain public records?

Queensland's *Public Records Act 2002* (the Act) takes a broad definition of 'public records.' A public record is any form of recorded information, either received or created by a public authority, which provides evidence of the business or affairs of that public authority. Based on this definition Queensland State Archives took the view that most, if not all, business systems within the scope of the *Queensland Government 2012 ICT Audit* would contain some public records. Any system containing copies of records (for example, where data had already been migrated to another system) were also viewed as holding public records, based on the express provision in the Act which states that a public record includes a copy of a public record.

Further, the Toolkit references Queensland State Archives' existing published advice on the topic *What is a public record in the digital environment* [Queensland State Archives 2013b] and to ISO 16175-3:2010 [International Standards Organisation 2010] to help identify the records.

3.1.2 How long do those records need to be kept?

For those records which are covered by an existing disposition authorisation, the retention periods are clearly defined. But for those records which have never been appraised, planning the preservation needs of the records in systems earmarked for decommissioning without an objective appraisal of the value of records would be difficult/almost impossible.

As Queensland State Archives had no published guidance on how to determine retention periods for records (though advice existed around justifying retention periods), a high level appraisal matrix needed to support this important step in the workflow.

The high-level appraisal advice guides public authorities through a simple appraisal exercise to determine (at a high level) how long those public records are **likely** to be required to meet business, legal, social, historical and other needs.

If this appraisal determines that the records are low value and could actually be past their potential 'use by date', agencies may seek a one-off approval to dispose of those records. Importantly, undertaking the high level appraisal does not waive the

[1] http://www.archives.qld.gov.au/Recordkeeping/BusinessSystems/DecomWorkflow/Documents/Full workflow diagram.pdf

requirement to seek the State Archivist's authorisation to dispose of the records.

Since the release of the Methodology and Toolkit, the Public Records Office of Victoria has undertaken its own study into the state of significant databases across the Victorian government. The final report highlights, among other things, the value of a high level appraisal tool such as the one developed by Queensland State Archives to help public authorities identify which system contain high value permanent and long term records as a first step to managing their legacy systems. In the absence of other published models, this particular tool may well have interest to other archives beyond Queensland's borders.

3.1.3 Separating temporary and permanent value records in systems for disposal or preservation

Despite many Queensland public authorities having a current disposition authorisation, some agencies continue to struggle to maintain effective control over their digital information through the proactive deletion of records. Anecdotally, Queensland State Archives is aware that there are several reasons why this is the case, including:

- many systems do not have disposal functionality enabling time-expired records to be removed from the system
- the quality of some record metadata is such that the task of matching records to record classes in disposition authorisations may be onerous.

The Toolkit advice on sentencing (selection) essentially empowers agencies to take a risk management approach: giving them an understanding of the issues and implications of sentencing public records at the individual record level or the system level. For practical reasons, the advice leans towards sentencing at the system level, as record level sentencing is generally more time consuming and may be impractical even if the system has the technical capabilities. In cases where, taking into account the longest retention period applicable to records in the system, the records are nearing expiry, the cost of preserving the entire system may be more cost effective than sentencing and disposing of portions of the data. This is a judgment which individual public authorities ultimately need to make, but their decision will hopefully be a more informed one.

3.1.4 What factors need to be weighed when determining the most appropriate preservation solution?

In the absence of digital preservation services or infrastructure or a comprehensive Queensland government preservation framework, Queensland State Archives approach to the issue of how best to ensure records in decommissioning candidate systems are preserved for as long as they are legally required has been to provide agencies with advice on a number of acceptable options. Each option has benefits and risks which need to be evaluated by a public authority, with appropriate mitigation strategies put in place to address all risks. The options presented are:

- Migrate the records and preserve them in a managed recordkeeping environment:
- Actively manage the records in the original business system by either virtualisation methods or retaining the system on the original software and hardware platform

Printing records to paper is addressed but discounted as the option of 'last resort' in answer to the suggestion frequently put forward by some agencies in discussions with Queensland State Archives on decommissioning issues, as an appropriate (i.e. financially practical) solution. Print to paper is not encouraged firstly because any record which can be printed to PDF can be retained in digital form without the need to print, and because a static representation of records designed to be used and viewed in a variety of ways inevitably reduces the completeness, usability, and authenticity of the original records.

4. IMPLEMENTATION

Between August and November 2014, a review is being undertaken to test whether the methodology achieved its intended outcomes. Key findings from this benefit review are expected to be finalized by the end of the year. Queensland State Archives will survey and interview all 20 state government departments to find out:

- The extent of uptake of the methodology
- Whether agencies apply the methodology to real systems
- Key areas for revision or additional guidance needed
- Learnings in relation to how the methodology was applied to different types of systems and records.

Interim results at time of writing show that:

- Agencies have decommissioned a number of systems using the methodology
- Agencies use a risk management approach in assessing the likelihood that business systems and databases contained high value long term records
- In most of these systems the records have been completely migrated to new business systems, and so the systems were decommissioned under the General Retention and Disposal Schedule for digital source records
- Many agencies see successive migration as a viable strategy to preserve high value long term records for as long as required.
- Agencies want the ability to transfer periodic snapshots of permanent value records held in agency business systems to a permanent whole-of-government digital archive
- Agencies wanted help to devise strategies to make the records accessible and to keep them accessible, especially with more complex formats such as GIS and business systems
- Records in some business systems have been exported in formats such as PDF or spreadsheets and stored in the agency electronic digital records management systems.
- Some records were exported and printed to paper
- The relational database layer in some business systems was exported to an SQL relational database management system with some stock queries and reports designed to answer common questions
- Most of these exported records were of a short term temporary nature, but others were at risk of loss through technological obsolescence
- Agencies used the toolkit to devise new policies, tools, check lists and templates to ensure that recordkeeping and disposal is considered during the design of replacement systems and

that the record migration methodology is adequately documented.

5. FOLLOW ON WORK

Queensland State Archives intends to improve the toolkit and include practical examples based on real-life implementations.

Queensland State Archives has undertaken to report periodically on the number of applications for disposition authorisation and notifications of lost digital records, to the Public Records Review Committee and Queensland's integrity agencies. However, to date no applications have been received.

Queensland State Archives will develop methodologies to allow agencies to identify, preserve and provide access to long term value records that remain in their custody after the business system that created them is decommissioned and no longer operational.[Fitzgerald 2013] The methodologies will be used where the business function has ceased and no replacement system exists, to provide periodic snapshots of records in an existing business system or in business systems being superseded and its records being migrated to a new system.

One possible methodology for relational database backed business systems involves mapping the archival records required to document system functions and transactions to the application screens and reports, identifying the corresponding SQL queries and adding these as views to the database layer before archiving with a database archiving tool such as the Swiss Federal Archives' SIARD tool. Preserving corresponding screen shots and report samples will enable agencies to reconstruct facsimiles of these from the archived data to provide more authentic and meaningful access when the business system is no longer available. The aim is to enable agencies to use this methodology to preserve and provide appropriate and meaningful access to long term value records in their custody or to create a Submission Information Package for transfer to a digital archive.

The ICT Audit highlighted the need for agencies to focus on the design of systems, to ensure that recordkeeping and disposal functionality is embedded in new business systems, processes and services from the outset, as it is difficult to resolve these matters effectively at the end of the life of the system. Queensland State Archives will build on the foundations of ISO 16175-3:2010: *Guidelines and functional requirements for records in business systems* [International Standards Organisation 2010] to produce policies, checklists, templates, tools, practical guidelines and case studies for specific business process types and include the policies, tools and templates developed by agencies to ensure new systems and services address these issues during design and implementation.

6. REFERENCES

[1] Fitzgerald, Neal. 2013. Using data archiving tools to preserve archival records in business systems In *Proceedings of the 10th International Conference on Preservation of Digital Objects* (Lisbon, Portugal, 02-06 September 2013)

[2] Gartner. 2011. "Information Management in the 21st century", (Sept 2011)

[3] International Standards Organisation. 2010. *ISO 16175-3:2010: Information and documentation – Principles and functional requirements for records in electronic office environments – Part 3: Guidelines and functional requirements for records in business systems* accessible from: http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=55792

[4] Queensland Government. 2012. *Queensland Government ICT Audit* (Oct 2012) www.qld.gov.au/dsitia/assets/documents/ict-audit-2012-report-a.pdf

[5] Queensland Government Chief Information Office. 2011. *Application Rationalisation Methodology* (Jun 2011). [Queensland Government only]

[6] Queensland State Archives. 2012. *General Retention and Disposal Schedule for digital source records* (QDAN678 v1). (Jun 2012) http://www.archives.qld.gov.au/Recordkeeping/BusinessSystems/Pages/Advice-toolkit.aspx

[7] Queensland State Archives. 2013. *Managing public records when decommissioning business systems methodology* and toolkit accessible from: http://www.archives.qld.gov.au/Recordkeeping/BusinessSystems/DecomWorkflow/Pages/workflow.aspx

[8] Queensland State Archives. 2013. *What is a public record in the digital environment* accessible from: http://www.archives.qld.gov.au/Recordkeeping/BusinessSystems/Pages/Advice-toolkit.aspx

A Digital Preservation Environment Maturity Matrix for NSLA Libraries

Sarah Slade
State Library of Victoria
328 Swanston Street,
Melbourne, VIC 3000, Australia
+61 (03) 8664 7383
sslade@slv.vic.gov.au

David Pearson
National Library of Australia
Parkes Place,
Parkes, ACT 2600, Australia
+61 (02) 6262 1570
dapearso@nla.gov.au

Libor Coufal
National Library of Australia
Parkes Place,
Parkes, ACT 2600, Australia
+61 (02) 6262 1831
lcoufal@nla.gov.au

ABSTRACT

In this paper, the authors describe the work done within the NSLA Digital Preservation Group to create a list and description of the functional components of an ideal digital preservation environment and a matrix of the current stage of development against each component for each NSLA library. After defining underlying assumptions, the functional components were derived from the OAIS standard. A modified Capability Maturity Model was incorporated as a mechanism for determining each organisation's stage of development against each component. The matrix was then completed by representatives from the Digital Preservation Group in each of the ten NSLA libraries. The respondents were asked to self-rate their organisations for both the current digital preservation situation, and an intended situation in three years' time. NSLA has identified digital preservation as an area of priority. The results from the Digital Preservation Environment Maturity Matrix reveal that NSLA libraries are on the right path but have some way to go before digital preservation processes are mature, sustainable and fit for purpose. Collaboration on policies, products and infrastructure will continue to address these needs.

General Terms

Management, Measurement.

Keywords

Digital Preservation, Capability Assessment, Environment Maturity Matrix, NSLA Libraries.

1. INTRODUCTION

In July 2012, the National and State Libraries of Australasia (NSLA) established a Digital Preservation Group. NSLA is comprised of the National Library of Australia, National Library of New Zealand, State Library of Victoria, State Library of New South Wales, State Library of Queensland, State Library of South Australia, State Library of Western Australia, Northern Territory Library, LINC Tasmania and Libraries ACT. The individual libraries are at differing states in their digital collecting maturity. They all are building and providing access to digital collections but only a few have active digital preservation systems and programs in place.

The objectives of the NSLA Digital Preservation Group were to:

- Gain a shared understanding of current digital collection management practices and workflows in NSLA libraries.
- Share information about digital preservation best practice.
- Determine the core requirements for managing the preservation of digital collections in NSLA libraries and identify opportunities for collaboration.

These objectives took into account the different stages of NSLA libraries in the adoption, development and implementation of digital preservation.

At the time the Group was established it identified six key work packages:

- 1: What is it and why? A statement on digital preservation and set of principles.
- 2: How well? A Digital Preservation Environment Maturity Matrix.
- 3: Who? A Digital Preservation Organisational Capability and Skills Maturity Matrix.
- 4: Nuts and Bolts: A common technical registry for NSLA libraries of file formats with software and hardware dependencies.
- 5: Collaboration and Partnership: A summary of opportunities for international representation and collaboration.
- 6: Confronting the Abyss: A business case for dedicated research into how to preserve difficult digital object types.

This paper focuses on work package 2, describing the work in creating the Digital Preservation Environment Maturity Matrix, and the initial findings from its use by the ten NSLA libraries.

The aim of the work package was to create a list and description of the functional components of an ideal digital preservation environment (keeping the list to a maximum of 20 components) and a matrix of the current stage of development against each component for each NSLA library [1] [2]. The benefits of this approach were designed:

- Firstly, to identify where various libraries currently sit against the list;
- Secondly, to help the libraries to identify development needs; and
- Thirdly, to help NSLA identify collaboration development needs.

The NSLA Digital Preservation Environment Maturity Matrix itself was developed by the Group over the first year of its operation and approved for use by the NSLA CEOs in March 2013. Representatives from the Digital Preservation Group in each of the ten NSLA libraries then completed the matrix (Refer to Table 1 in appendix A), and a summary report was written on these initial findings; outlining the current level of digital preservation maturity across NSLA.

1.1 Related Work

It should be noted that similar work is currently being conducted in at least two other international projects:

- The National Digital Stewardship Alliance in the USA has developed Levels of Digital Preservation, a tiered set of recommendations for how organisations should begin to build or enhance their digital preservation activities. This was used to assess the current state of digital preservation among the NDSA members [3] [4].
- BenchmarkDP, a three-year research project funded by the Vienna Science and Technology Fund, is developing a coherent, systematic approach to assess and compare digital preservation processes, systems and organisational capabilities [5] [6].

The NSLA Digital Preservation Group, via the National Library of Australia, has been in touch with the former and participated in a survey and an iPres workshop organised by the latter. The Group will follow their progress to identify potential areas which could feed into the matrix in the future.

2. MATRIX DEVELOPMENT

The development of the NSLA Digital Preservation Environment Maturity Matrix fell into key areas:

1. Confirmation of the underlying assumptions
2. Identification of the functional components
3. Use of a maturity model

2.1 Underlying Assumptions

The first stage in the development of the matrix was to define a set of underlying assumptions on which the functional components are based. These assumptions were provided at the start of the matrix.

Interestingly, some members of the Group indicated that the assumptions were not necessarily valid for them, which would have made it difficult to evaluate their response. Therefore, each respondent was asked to state whether the assumption was correct for their organisation at the time of completion of the matrix. This ensured transparency around the assumptions each organisation would be making when completing the matrix, providing an additional level of confidence when comparing the results over time as the completion of the matrix is repeated.

The underlying assumptions in the matrix are that an organisation:

- Is actively collecting digital materials, both born digital and digitised.
- Is committed to preserving its digital materials for the long term.
- Has resources (including staff or vendor with appropriate skills) dedicated to the task.
- Has a sustainable funding model.
- Aims to comply with the open archival information system (OAIS) responsibilities as listed in the matrix.

2.2 Functional Components

Once the underlying assumptions had been defined the next step in the development of the matrix was to identify the functional components.

An ideal digital preservation environment should contain a mix of policies, processes and resources (including staff and technologies). The reference model for an open archival information system (OAIS) is a commonly accepted standard among the digital preservation community [7]. The OAIS standard is a high-level, abstract model, which, amongst other things, "provides a framework, including terminology and concepts, for describing and comparing architectures and operations of existing and future archives". This makes it a good starting point for describing the high-level functional components of an ideal digital preservation environment. A similar approach has been taken previously by a JISC funded project involving The National Archives (TNA) and the UK Data Archive at the University of Essex between 2004 and 2005 [8].

In order to be compliant with the OAIS standard, each library needs to address the following responsibilities:

- Negotiate for and accept information from information producers.
- Obtain sufficient control of the information for long-term preservation.
- Determine the designated user community.
- Ensure the information is independently understandable to the designated community without the need of special resources.
- Follow documented preservation policies and procedures, which ensure that the information is preserved against all reasonable contingencies.
- Make the information available to the designated community [7].

Rather than simply listing the functional components, a set of generic and open-ended questions were framed, which were based on selected functions of individual OAIS entities (refer to Table 2).

The questions were intended to act as a guide and help respondents identify and describe their organisation's current level of digital preservation maturity, as well as assist in planning for the future.

This opened the way for each organisation to determine how to approach the challenges of digital preservation in a manner that best suited their needs. It also acknowledged that an "ideal" digital preservation environment is still to be defined. In applying the framework defined in the matrix in this way it was hoped that the functional components for a digital preservation system could be inferred from the questions and the institutional responses.

The top-level headings of the list followed the functional entities of the OAIS model. The individual questions under each heading were based on selected functions of individual OAIS entities, with some modifications made to the selected functions for the purpose of simplicity and clarity.

Table 2 High level functional components of a digital preservation environment

1. Pre-ingest Activities
What system policies and standards related to digital collecting do you have in place in your library?
What system policies and standards related to digital preservation do you have in place in your library?
2. Ingest
What SIPs do you receive from producers, and how?
How do you validate the SIPs?
How do you generate AIPs from SIPs?
What metadata do you extract from AIPs or collect from other sources, and how?
3. Archival Storage
How are your AIPs stored?
What proactive measures do you take to refresh your archival media/storage?
What routine and special error checking do you perform to make sure that no components of the AIP are corrupted in archival storage or during any internal archival storage data transfers?
What IT disaster recovery plans and business continuity plans does your library have in place to protect your digital assets?
4. Data Management
How do you store, maintain and update metadata for your library's digital collection content?
How do you monitor collection status?

5. Administration
How do you negotiate submission agreements and audit submissions to ensure that they meet your institution's standards?
How do you manage system configuration?
What mechanisms do you provide to restrict or allow physical access to elements of the archive, as determined by archive policies?
How do you establish and maintain system standards and policies?
6. Digital Preservation Planning
How do you monitor changes in the Digital Preservation and ICT technology environments and in the designated community's service requirements and knowledge base?
How do you develop preservation strategies and standards?
How do you develop packaging designs and preservation action plans?
7. Access
How do you provide access to your data?
How do you ensure that the user is authorised to access and receive the requested items?

2.3 Maturity Model

The final stage in the development of the matrix was to incorporate a mechanism for determining each organisation's stage of development against each component. To achieve this, the Group modified the Capability Maturity Model (CMM) [9].

Although CMM was originally developed to measure and manage the improvement in software development processes, the model is flexible and adaptable to more diverse subject areas, such as digital preservation.

There are five levels in the CMM defined as:

Level 1 - Initial

Level 2 - Repeatable

Level 3 - Defined

Level 4 - Managed

Level 5 - Optimising

Detailed definitions were provided with the matrix as examples to demonstrate how CMM could be adapted for use across NSLA to assess the level of digital preservation activities currently in place. These definitions are summarised below:

Level 1 – Initial

At level 1 maturity:

- Processes are usually ad hoc.
- Achievement depends on the competence of the people in the organisation and not on the use of proven processes.
- Organisations often produce products and services that work, but frequently exceed both budget and schedule.

Level 2 – Repeatable

At level 2 maturity:

- Digital preservation achievements are repeatable, but the processes may not repeat for all digital preservation activities in the organisation.
- Process discipline helps ensure that existing practices are retained during times of high pressure.
- Basic digital preservation processes are established to track cost and to match activities to agreed digital preservation objectives.
- There is still a significant risk of exceeding cost and time estimates for the identified activities.

Level 3 - Defined

In addition to meeting the activities in level 2, at level 3 maturity:

- Digital preservation activities are performed and managed according to documented plans.
- The status and the delivery of digital preservation activities and services are visible to management at defined points.
- Standard organisational processes for digital preservation are established and improved over time.
- These standard processes are used to establish consistency across the organisation.
- Management defines digital preservation objectives and ensures that these objectives are met.

Level 4 – Managed

At level 4 maturity:

- Management can effectively control the digital preservation effort, using precise measurements.
- In particular, management can identify ways to adjust and adapt the digital preservation effort to particular activities without measurable losses of quality or deviations from specifications.
- The organisation sets a quantitative quality goal for both digital preservation process and ongoing maintenance and support.
- Sub-processes are selected that significantly contribute to overall performance and the selected sub-processes are controlled using statistical and other quantitative techniques.

Level 5 - Optimising

- At this level the organisation focuses on continually improving process performance through both incremental and innovative technological improvements.
- Quantitative process-improvement objectives are established, continually revised to reflect changing business objectives, and used as criteria in managing process improvement.
- The effects of deployed digital preservation process improvements are measured and evaluated against the quantitative process-improvement objectives.
- Both the defined processes and the organisation's set of standard digital preservation activities and processes are targets of measurable improvement activities.
- Optimising processes that are nimble, adaptable and innovative depends on the participation of an empowered workforce aligned with the business values and objectives of the organisation.
- The organisation's ability to rapidly respond to changes and opportunities is enhanced by finding ways to accelerate and share learning.

Using this five level rating system, in the matrix respondents were asked to self-rate their organisations for both the current digital preservation situation, and an intended situation in three years' time. This allowed an organisation's digital preservation aims to be captured, as well as their current level of activity. As the matrix will be repeated over time it will also allow their actual achievements against these aims to be compared.

As well as the columns for self-rating their level of current and intended maturity against each functional component, the matrix also included a column for commenting on the current state of digital preservation with the institution, providing the opportunity for additional context to be provided to the maturity ratings.

3. MATRIX USE

The completed matrix was distributed to the ten NSLA libraries in February 2013. All initial submissions were received by 22 August 2013. These were discussed at the NSLA Digital Preservation Group meeting in Adelaide, Australia, in September 2013.

At that meeting it was decided to make minor changes to the matrix to ensure a consistent approach to the responses, and NSLA libraries were able to review and modify their responses as required. All final responses were received in October 2013 and integrated into a final report which was signed off by the NSLA CEOs in November 2013 [2].

3.1 Analysis of the Initial Results

The analysis of the responses focused on the assumptions and CMM ratings. The respondents' comments were made available to the NSLA libraries but were not analysed. The overall picture revealed by the matrix across NSLA libraries has been included in this paper, without identification of individual libraries.

3.2 Underlying Assumptions

All of the ten NSLA libraries completed the matrix and reported the following underlying assumptions:

- All libraries are collecting digital materials.
- All libraries are committed to preserving access to their content over time.
- Six out of the ten libraries did not have resources (including staff or vendor with appropriate skills) dedicated to the task of digital preservation.
- Eight out of the ten libraries did not have a sustainable funding model for digital preservation.
- All libraries aim to comply with OAIS responsibilities.

3.3 Matrix Responses

The responses to the matrix were then analysed, as summarised in tables 3 and 4 in appendices B and C.

Overall, they demonstrated a clear picture of the state of digital preservation within NSLA libraries, and these results were felt to be valid and useful. However, it was difficult to compare the results between the NSLA libraries, in part because the way the maturity model was applied may have led to subjectiveness in the self-assessment.

The responses to questions about the current state of digital preservation with the NSLA libraries, as detailed in table 3 in appendix B, revealed that:

- All NSLA libraries rated themselves well for providing and authorising access to digital collection material, both internally and externally. This also included managing and controlling physical access.
- All NSLA libraries appeared to be doing reasonably well on policies, but more so for collecting than preservation.
- Importantly, the rating for storage of digital materials seemed to be quite low and most NSLA libraries were not actively managing bit-level preservation. Although on average, refreshing media/storage and IT disaster planning seemed to be better managed, the figures are still a concern for some of the smaller libraries.
- All NSLA libraries rated themselves low for digital preservation planning, which shows that they are not yet in a position to do active preservation.

In the rest of the areas, it proved difficult to draw any concrete conclusion because there were large variations between the results for individual libraries. It should be noted that this might have been caused by the unavoidable subjectiveness of the assessments as stated above.

The responses to questions about the intended future state of digital preservation within the NSLA libraries, as detailed in table 4 in appendix C, revealed that:

- There is a large variation in the plans as the ratings for all but two questions range from 1 to 5.
- A small number of NSLA libraries indicated that they did not plan to improve their processes and some planned to stay at an ad hoc level. However, there was an agreement between the majority of libraries that they would like to (sometimes quite significantly) improve their current processes.
- For all questions, over half of the libraries would like to score at least 3, with some indicating that they aim to achieve 5 in three years' time. This was even higher for the last two question regarding Access with over half of the libraries ranking themselves at least 4.

4. FINDINGS FROM THE WORK AND OTHER STUDIES

In general the submissions demonstrated some issues with the questionnaire/matrix:

- The questions in the survey were rather open ended in order not to prescribe answers which may have potentially caused problems in the answers to certain questions. As pointed out by Kulovits cited in [6] 'a clear distinction between business process and information system' is needed.
- The initial analysis also demonstrated that the OAIS reference model on which the survey is based is a very complex concept which made answering the questions challenging for libraries that did not have a detailed understanding of the model**.**
- The CMM methodology proved to be difficult to apply consistently and to achieve objective results for:
    - Within the questionnaire.
    - Applied between libraries.

  However, the generalisations that are provided in 3.3 are felt to be valid and useful.
- Based on the above, some of the assessments raise concerns about whether the results can be taken at face value.

It also must be stressed that CMM is by no means the only approach that could be adopted by the NSLA Digital Preservation Group for assessing the level of digital preservation activities across the NSLA organization. Katuu, as cited in [6] and the Australian National Data Service [10] [11], provide other potentially useful examples.

5. NEXT STEPS

The NSLA Digital Preservation Group intends to continue to develop and extend the environment matrix. The Group may also consider analysing the detailed comments provided with each library's response further.

The Group had also developed a sister matrix, the Digital Preservation Organisational Capability Maturity Matrix (Work Package 3: Who?) that examines how well management and human resource practices support the evolution of digital preservation needs within NSLA libraries. The matrix has been completed and an initial analysis of these findings has been undertaken. At the time of writing, these findings are still to be discussed at a meeting of the Digital Preservation Group.

The NSLA Digital Preservation Group is planning to investigate integrating the results of the environment maturity and the organisational capability maturity matrices to provide a clear picture of progress or inactivity in the area of digital preservation for NSLA and individual libraries.

In addition, in November 2013 the NSLA Digital Preservation Group and ADRI (Australasian Digital Recordkeeping Initiative) met and discussed potential mutual initiatives. At that meeting it was decided that the Archives sector, as represented by ADRI, would also fill out the environment maturity matrix. At the time of writing ADRI members were in the process of doing this and the results are yet to be analysed or compared to those from the NSLA libraries.

The development of a combined environment maturity and organisational capability maturity matrix, combined with the ADRI results would provide a more holistic picture of the state of Digital Preservation in these two sectors in Australasia.

6. CONCLUSION

Overall this work has demonstrated the current variable maturity of NSLA libraries to deal effectively with the preservation of digital materials in their custody. Although some NSLA libraries are more mature than others in some aspects, all libraries are relatively immature in digital preservation matters. This was to be expected for logical preservation but it is unexpected that it is also the case for bit-level preservation (fixity checking, backups, storage media refreshing etc.). Without the preservation of the bits, the ability to preserve the logical content of the files over time is seriously compromised.

All libraries (at varying levels) indicated that they require sustainable funding and staffing models. The survey also demonstrated a need to develop or improve their capability through scalable ways to ingest digital content, collect technical metadata as well as monitor, plan and take preservation actions over time.

NSLA has identified digital preservation as an area of priority. The importance of this area to NSLA libraries is reflected in the creation of the Digital Preservation Group and its support of the Group's work to date. The results from the Digital Preservation Environment Maturity Matrix reveal that NSLA libraries are on the right path but have some way to go before digital preservation processes are mature, sustainable and fit for purpose. Collaboration on policies, products and infrastructure will continue to address these needs.

7. ACKNOWLEDGMENTS

The Authors would like to thank Geoff Hinchcliffe and Scott Wajon (SLNSW) and Shauna Kelly (NLA) for their input into the initial questions and final report. Thanks also go to Steve Knight (NLNZ) for his support for this paper. We would also like to thank NSLA and the individual NSLA libraries for their involvement in this work.

8. REFERENCES

[1] NSLA Digital Preservation: Project Scope, deliverable 2. Available at http://www.nsla.org.au/sites/default/files/publications/3.2 Digital Preservation.pdf

[2] Pearson, D. and Coufal, L. 2013. National and State Libraries of Australasia, Digital Preservation Work Package 2: Ideal Digital Preservation environment and a matrix of the current stage of development against each component for each NSLA library. Unpublished NSLA Report, Melbourne.
For a copy of this report contact the authors.

[3] NSDA Levels of Preservation. Available at http://www.digitalpreservation.gov/ndsa/activities/levels.html

[4] Phillips, M., Bailey, J., Goethals, A. and Owens, T. 2013. The NDSA Levels of Digital Preservation: An Explanation and Uses. IS&T Archiving, Washington. http://www.digitalpreservation.gov/ndsa/working_groups/documents/NDSA_Levels_Archiving_2013.pdf

[5] BenchmarkDP. Available at http://benchmark-dp.org/

[6] Becker, C. and Cardoso, E. 2014. 'Report on the Capability Assessment and Improvement Workshop (CAIW) at iPRES 2013' in *D-Lib Magazine*, Vol. 20, No. 3/4 (March/April 2014).

[7] OAIS ISO 14721:2012. Available at http://public.ccsds.org/publications/archive/650x0m2.pdf

[8] Beedham, H., Missen, J., Palmer, M. and Ruusalepp, R. 2005. 'Assessment of UKDA and TNA Compliance with OAIS and METS Standards'. Available at http://www.jisc.ac.uk/media/documents/programmes/preservation/oaismets.pdf

[9] Capability Maturity Model (CMM Self Rating). Available at http://www.selectbs.com/process-maturity/what-is-the-capability-maturity-model

[10] ANDS Research Data Management Framework: Capability Maturity Guide. Available at http://ands.org.au/guides/dmframework/dmf-capability-maturity-guide.html

[11] ANDS Research Data Management Framework: Capability Maturity Guide. Available at http://ands.org.au/assets/images/guides/dmf-capability-maturity-guide.png

Appendix A

Table 1 NSLA Libraries completing the matrix

NSLA Library	Location	Ongoing staff (2013) **Source:** *NSLA Workforce Data Report, November 2013*	Matrix completed by
Libraries ACT	Canberra	n/a	Senior Management staff
National Library of Australia	Canberra	418	Digital Preservation staff
National Library of New Zealand	Wellington	299	Digital Preservation staff
State Library of New South Wales	Sydney	300	Senior Management staff
Northern Territory Library	Darwin	53 (2012)	Senior Management staff
State Library of Queensland	Brisbane	233	Physical Preservation staff
State Library of South Australia	Adelaide	131	Senior Management staff
LINC Tasmania	Hobart	350	Senior Management staff
State Library of Victoria	Melbourne	286	Technology & Collection staff
State Library of Western Australia	Perth	170	Senior Management staff

Appendix B

Table 3 Statistical analysis of the current CMM ratings

Current	CMM Rating Count					Mean	Median	Mode
	1	2	3	4	5			
1. Pre-ingest activities								
What system policies and standards related to digital collecting do you have in place in your Library?	4	4	1	1	0	1.9	2	-
What system policies and standards related to digital preservation do you have in place in your Library?	6	3	1	0	0	1.5	1	1
2. Ingest								
What SIPs do you receive from producers, and how?	6	2	2	0	0	1.6	1	1
How do you validate the SIPs?	6	3	0	1	0	1.6	1	1
How do you generate AIPs from SIPs?	5	3	0	2	0	1.9	1-2	1
What metadata do you extract from AIPs or collect from other sources, and how?	6	3	0	1	0	1.6	1-2	1
3. Archival storage								
How are your AIPs stored?	5	3	1	1	0	1.8	1-2	1
What proactive measures do you take to refresh your archival media/storage?	5	3	0	2	0	1.9	1-2	1
What routine and special error checking do you perform to make sure that no components of the AIP are corrupted in archival storage or during any internal archival storage data transfers?	6	1	3	0	0	1.7	1	1
What IT disaster recovery plans and business continuity plans does your Library have in place to protect your digital assets?	3	4	2	1	0	2.1	2	2
4. Data management								
How do you store, maintain and update metadata for your Library's digital collection content?	2	6	1	1	0	2.1	2	2
How do you monitor collection status?	7	2	1	0	0	1.4	1	1
5. Administration								
How do you negotiate submission agreements and audit submissions to ensure that they meet your institution's standards?	5	2	3	0	0	1.8	1-2	1
How do you manage system configuration?	5	2	2	1	0	1.9	1-2	1
What mechanisms do you provide to restrict or allow physical access to elements of the archive, as determined by archive policies?	2	1	4	3	0	2.8	3	3
How do you establish and maintain system standards and policies?	4	2	4	0	0	2.0	2	-
6. Digital preservation planning								
How do you monitor changes in the Digital Preservation and ICT technology environments and in the designated community's service requirements and knowledge base?	7	2	1	0	0	1.4	1	1
How do you develop preservation strategies and standards?	7	1	2	0	0	1.5	1	1
How do you develop packaging designs and preservation actions plans?	7	2	1	0	0	1.4	1	1
7. Access								
How do you provide access to your data?	1	1	5	3	0	3.0	3	3
How do you assure that the user is authorised to access and receive the requested items?	1	1	7	1	0	2.8	3	3

Appendix C

Table 4 Statistical analysis of the future CMM ratings

Future	CMM Rating Count					Mean	Median	Mode
	1	2	3	4	5			
1. Pre-ingest activities								
What system policies and standards related to digital collecting do you have in place in your Library?	2	2	1	3	2	3.1	3-4	4
What system policies and standards related to digital preservation do you have in place in your Library?	3	1	2	3	1	2.8	3	-
2. Ingest								
What SIPs do you receive from producers, and how?	2	3	2	2	1	2.7	2-3	2
How do you validate the SIPs?	2	3	1	3	1	2.8	2-3	-
How do you generate AIPs from SIPs?	1	3	2	2	2	3.1	3	2
What metadata do you extract from AIPs or collect from other sources, and how?	1	3	2	3	1	3	3	2
3. Archival storage								
How are your AIPs stored?	1	4	1	2	2	3	2-3	2
What proactive measures do you take to refresh your archival media/storage?	1	4	0	2	3	3.2	4	2
What routine and special error checking do you perform to make sure that no components of the AIP are corrupted in archival storage or during any internal archival storage data transfers?	1	3	1	2	3	3.3	3-4	-
What IT disaster recovery plans and business continuity plans does your Library have in place to protect your digital assets?	0	4	1	1	4	3.5	3-4	-
4. Data management								
How do you store, maintain and update metadata for your Library's digital collection content?	0	4	2	3	1	3.1	3	2
How do you monitor collection status?	2	3	2	2	1	2.7	2-3	2
5. Administration								
How do you negotiate submission agreements and audit submissions to ensure that they meet your institution's standards?	2	3	2	2	1	2.7	2-3	2
How do you manage system configuration?	2	2	3	2	1	2.8	3	3
What mechanisms do you provide to restrict or allow physical access to elements of the archive, as determined by archive policies?	1	2	1	5	1	3.3	4	4
How do you establish and maintain system standards and policies?	2	1	2	3	2	3.2	3-4	4
6. Digital preservation planning								
How do you monitor changes in the Digital Preservation and ICT technology environments and in the designated community's service requirements and knowledge base?	2	1	4	2	1	2.9	3	3
How do you develop preservation strategies and standards?	1	2	4	2	1	3	3	3
How do you develop packaging designs and preservation actions plans?	2	2	3	2	1	2.8	3	3
7. Access								
How do you provide access to your data?	1	0	1	5	3	3.9	4	4
How do you assure that the user is authorised to access and receive the requested items?	1	1	1	4	3	3.7	4	4

Panels

Getting to Digital Preservation Tools that "Just Work"

Andrea Goethals
Harvard Library
90 Mt. Auburn St.
Cambridge, MA 02138
+1 (617) 495-3724
andrea_goethals@harvard.edu

Paul Wheatley
Paul Wheatley Consulting Ltd
Leeds, LS2 9BS
paulrobertwheatley@gmail.com

Stephen Abrams
UC Curation Center, CDL
415 20th Street, 4th Floor
Oakland, CA 94612
+1 (510) 987-0425
Stephen.Abrams@ucop.edu

Janet Delve
University of Portsmouth
Eldon Building
Winston Churchill Avenue
PO1 2DJ
+44 (0)23 9284 5524
janet.delve@port.ac.uk

Ed Fay
Open Planets Foundation
c/o The British Library, Boston Spa
Wetherby, West Yorkshire, LS23 7BQ
+44 1937 54 6013
ed@openplanetsfoundation.org

Cal Lee
UNC Chapel Hill
Manning Hall, Room 212
Chapel Hill, NC 27599
+1 (919) 962-8071
callee@ils.unc.edu

Dirk von Suchodoletz
Faculty of Engineering
Albert-Ludwigs University Freiburg
Freiburg i. B., Germany
dirk.von.suchodoletz@rz.uni-freiburg.de

ABSTRACT

In this paper, we describe a panel that discussed experiences, strategies, challenges and lessons learned maintaining and funding digital preservation tools that are available for use by the digital preservation community. The panel, together with the audience, explored the challenges and discussed potential solutions to developing more robust and sustainable support structures for these tools.

General Terms

Infrastructure, Communities, Digital preservation marketplace

Keywords

digital preservation tools, open-source, enhancements, software development, funding

1. INTRODUCTION

Many of the tools our institutions rely on for digital preservation planning and activities are maintained and funded by single institutions, or reliant on short-term funding. In some cases these tools had project or grant funding that has run out. This presents challenges for the maintaining institutions, funding agencies where applicable, and the digital preservation community as a whole which is reliant on these tools.

- In these cases the maintaining organizations have made their tools available to the community for use but often are not able to keep up with the growing and changing needs of the larger community. Even if the tools are open sourced, the maintaining institutions typically do not have the resources to test, incorporate and document the contributed changes in a timely way.
- Funding agencies and governments want the products they fund to remain relevant and usable and have broad impact well beyond the project funding period. They want to know that any preservation tools they fund can be sustained and improved over time.
- Organizations using the tools can get frustrated when the tools do not improve at the rate they would like and enhancements that they would like are not added. These organizations are not able to adequately improve their preservation practices and infrastructure in a timely way without having reliable tools that meet their requirements.

2. THE TOOLS

The panelists represented a variety of tools:

- BitCurator
- Emulation as a Service (EaaS) Project Tools
- File Information Tool Set (FITS)
- JHOVE
- JHOVE2
- KEEP Project Emulation Tools
- PLANETS Project Tools
- SCAPE Project Tools
- Unified Digital Format Registry (UDFR)

The panelists briefly described the purpose and status of the tools and the key challenges they have faced enhancing, funding, "mainstreaming", governing and providing roadmaps for these tools. In some cases they told success stories where they had managed to forge sustainable models. The audience was invited to pose questions for the panelists and to contribute ideas for how these tools could be more easily improved and sustained.

3. DISCUSSION

The main points made by the panelists and audience members are summarized here.

3.1 Observations & Comments

- It was asserted that a very large amount of money has been spent on digital preservation tools that have resulted only in demonstrations and prototypes, and not usable code or well-used tools. Several people objected to this statement saying that there are many examples of tools developed within the digital preservation community that may not be perfect but that are widely used.

- The question of whether or not development within memory institutions is a good idea was raised. One responder said that in-house development is done because it is convenient, expedient or existing tools do not necessarily meet local requirements. While this can produce quick results it can pose quality and sustainability problems because of poor quality code, the need to maintain the code base long-term, etc. Others responded that it can be misleading to assume that in-house development within memory institutions is always done by librarians when it may be done by software developers and computer scientists as would be the case when developed by a commercial company.

- It was posited that developers and project managers could be trained in a month to produce good code adhering to best practices but this was disputed by some in the audience.

- The digital preservation community needs to become more proficient at hosting open source tools. While the open source approach for software development is favored by many institutions, some parts of it are not fully embraced because of a lack of resources to perform tasks such as code cleanup, adherence to good coding style, fixing reported bugs, and testing of patches.

- Management within organizations needs to be convinced to spend not only initial development resources on tool development, but also the ongoing maintenance costs which can be orders of magnitude more costly.

3.2 Success Stories

- There is widespread usage of some of these tools (e.g. JHOVE), both in stand-alone mode as well as integrated into larger repository systems.

- Some of these tools (e.g. BitCurator and the SCAPE tools) have found hosting environments (Educopia, OPF) and communities of use after their project funding ended.

3.3 Lessons Learned

- If you make the code available on an open source hosting platform like github be prepared for forks unless you have a clear documented process for developers to contribute code that can be easily integrated into the main branch.

- The key lessons learned from the UDFR registry are the importance of continued synchronization between registries such as PRONOM and UDFR to prevent the immediate divergence of format information, the role of the community and governance to sustain the registry, and dedicated evangelism to maintain interest and use of the registry.

- It can be hard to transition research projects into production tools. The focus of research is on new ideas, and the end products are seen to be publications and advanced degrees, and not necessarily the tools themselves. Often the tools remain prototypes which lack the documentation and commitment to ongoing maintenance that is needed for these tools to be used in production. There is little incentive for researchers to continue to work on tools that no longer are groundbreaking. This is a hard problem because it is appropriate for digital preservation research such as emulation to be conducted by academic or research institutions but for the reasons described it is difficult to transition this research into usable tools.

- Use cases should be well understood before technical development begins.

- Hackfests by design do not contribute to sustainable tools. They have been better at innovating new prototypes than enhancing existing production tools.

3.4 Ideas to Explore

- Up to now libraries and archives have carried the research and development cost for digital preservation. Can we find ways for other institutions (banking, etc.) with more money to contribute?

- Consider contracting development out to commercial companies. An example was given where commercial companies are asked to respond to an RFP developed by a group of memory institutions.

- Solicit and communicate stories where digital preservation tools have been found to be useful to institutions so that the stories can be shared with funders of the tools to encourage continued funding.

- Follow up after Hackfests with attention to clean up (documentation, testing, etc.).

- The digital preservation community could make a statement about adopting software development and testing best practices and funders could mandate that they be followed.

Preserving Government Business Systems

Cassie Findlay[1]
Project Manager, Digital Archives
State Records Authority of New South Wales
PO Box 516, Kingswood, NSW 2747, AUSTRALIA
+61 2 8805 5313
Cassandra.Findlay@records.nsw.gov.au

ABSTRACT

The role of government archival institutions is to ensure that essential evidence of the business of government is made, kept and used. This evidence now resides in a wide array of systems and structures; from large centralised case management systems to collaborative workspaces in the cloud. Government archives work with agencies on systems design, improvements and migrations with good recordkeeping their goal.

General Terms

infrastructure, communities, strategic environment, preservation strategies and workflows, case studies and best practice.

Keywords

Government archives, recordkeeping

1. THEME

The role of government archival institutions is to ensure that essential evidence of the business of government is made, kept and used. This evidence now resides in a wide array of systems and structures; from large centralised case management systems to collaborative workspaces in the cloud. Government archives work with agencies on systems design, improvements and migrations with good recordkeeping their goal. A critical part of this work is ensuring that digital records in complex business systems requiring permanent retention as archives are as well preserved and available as other records; a formidable challenge, to which solutions are still evolving. The need to retain meaning, authenticity and evidential qualities as well as usability must be balanced with the practical constraints of maintaining multiple systems in a single environment and with limited resources. Hear from three government archives on the approaches they are taking to the preservation of government business systems and how this work relates to digital preservation initiatives in other sectors.

2. PROGRAM

Time	Subject	Presenter
15 mins	Introduction: Government archives and the preservation challenge	Cassie Findlay
45 mins	Panel discussion Each panel member will describe their own experience in the preservation of business systems followed by a facilitated discussion. • What have been the challenges facing government archives in the preservation of digital business systems? • How do archival approaches differ from other sectors? How are they similar? • What approaches to the preservation of business systems show most promise for the future?	Panelists Andrew Waugh Richard Lehane Neal Fitzgerald
30 mins	Audience Q&A with the panel	Moderated by Cassie Findlay

3. PRESENTERS AND PANELISTS

Cassie Findlay is a recordkeeping consultant who was until recently the Project Manager, Digital Archives at the State Records Authority of NSW (Sydney, Australia). In this role she led a team responsible for the development and implementation of the NSW Government's first digital archive, dealing with a variety of government business systems along the way. She has a BA in history from the University of Sydney and a Master of

[1] Neal Fitzgerald, Principal Digital Archives Technology Analyst Queensland State Archives, PO Box 1397, Sunnybank Hills, QLD, 4109, AUSTRALIA, +61 7 3131 7982 (ext 87982), neal.fitzgerald@archives.qld.gov.au

Andrew Waugh, Senior Manager, Standards and Policy, Public Record Office Victoria, PO Box 2011, North Melbourne VIC 3051, AUSTRALIA, +61 3 9348 5724, andrew.waugh@prov.vic.gov.au

Dr Richard Lehane, Project Officer, Digital Archives, State Records Authority of New South Wales, PO Box 516, Kingswood, NSW 2747, AUSTRALIA, +61 2 9673 1788, richard.lehane@records.nsw.gov.au, Information Management (Records / Archives) from the University of New South Wales. Cassie has served on the National Council of the Australian Society of Archivists and is currently Project Lead on the review of the international standard for records management, ISO 15489. She is a co-founder of The Recordkeeping Roundtable (rkroundtable.org) and tweets as @CassPF.

Neal Fitzgerald is a senior technology research analyst in the Digital Archives unit at Queensland State Archives. Neal has worked at the State Library of Queensland configuring and supporting the applications managing the digital image, audio and video archives. Before that Neal worked as a database consultant to the corporate, government and community sectors. Neal has also worked for software companies and hardware vendors as a database specialist developing and supporting business systems and in the IT department at UTS in Sydney teaching database and information systems.

Andrew Waugh has been involved with digital preservation since 1998 when he was part of a team from the Australian research organisation CSIRO that worked with Public Record Office Victoria to develop the Victorian Electronic Records Strategy (VERS). Andrew was then heavily involved with the pilot implementation of VERS in a Victorian agency in 2001, and in 2002 was seconded to PROV to work in the VERS Centre of Excellence. During this time he was involved in the implementation of the PROV digital archive, as well as building tools to assist agencies in transferring digital records. Andrew is currently the Senior Manager, Standards and Policy, at Public Record Office Victoria (PROV) where he is responsible for the development of the standards and policies that govern recordkeeping within the Victorian government. Andrew has an MSc in Computer Science from the University of Melbourne and prior to coming to PROV was a scientist at the Australian research organisation CSIRO where he specialised in computer networking, metadata, resource discovery, and document management.

Richard Lehane is an archivist at the State Records Authority of NSW. He is a member of the digital archives team, who are undertaking a three year project to build a whole of government digital archive for New South Wales. Richard also works on State Records' Open Data project, <http://data.records.nsw.gov.au>, and new search engine, "Search" <http://search.records.nsw.gov.au>.

Digital Preservation: Are We Succeeding? Panel Debate

ABSTRACT

The Programme Committee created a panel for this year's conference that was structured to generate introspection in the digital preservation community. The panel took the form of a debate between international figures and debated the question: Digital Preservation: Are we succeeding? The following are notes compiled from the event, rather than a summary and conclusion. A video of the event is available on the iPRES 2014 website.

General Terms

infrastructure, communities, strategic environment, preservation strategies and workflows, specialist content types, digital preservation marketplace, theory of digital preservation, case studies and best practice, and training and education.

Keywords

Evaluation, success, failure, digital preservation.

1. INTRODUCTION

The Programme Committee created a panel for this year's conference that was structured to generate introspection in the digital preservation community. The panel took the form of a debate between international figures and debated the question: Digital Preservation: Are we succeeding?

The participants were assigned to one side of the discussion: pro or con. Their arguments did not necessarily reflect their own professional stance on the question, but rather were designed to provoke the audience to consider their own stance on the question.

1.1 Debaters

The following were the debaters for the session.

Moderator

Shaun Hendy, *University of Auckland*

Pro

Ross Wilkinson, *Australian National Data Service*

Helen Tibbo, *University of North Carolina at Chapel Hill*

Andi Rauber, *Technical University of Vienna*

Con

Seamus Ross, *University of Toronto*

Barbara Sierman, *National Library of the Netherlands*

Ed Fay, *Open Preservation Foundation*

1.2 Debate Agenda

As outlined in the agenda below, each debater was given three minutes to highlight the key arguments for their side. The floor was then opened up to the audience to ask questions of the debaters. Finally, the debaters were asked to list three things that they would like to see happen to either keep digital preservation succeeding, or to put it onto a path towards succeeding.

- Introductions
- Debate (3 mins each participant)
- Q&A from the audience
- Decision
- Top 3 things to help digital preservation follow a positive path
- Wrap up and final floor comments

1.3 Debate points

1.3.1 *Andi Rauber*

- Measurement of success – knowledge
  - We know much more about the issues
  - Solutions – big market for range of solutions
  - Jobs – creating new jobs – digital curation, data management, etc.
  - E-government another movement forward – we can guarantee persistence.

1.3.2 *Ed Fay*

- Where's the transparency across systems
- User experience – tools that don't just work
- No alignment with industry
- Siloed/isolated teams – still seen as fringe activity
- Missing the value argument – too abstract (preservation)
- Cross-domain requirements analysis
- Collaboration issues
- Enormous advocacy is needed

1.3.3 *Helen Tibbo*

- DP/DC as a field within 20 years (1995 start)
- Internet has been a key part of collaboration – not as much isolation – taking less time to develop
- Educational programs – graduate education, workshops, etc.
- Conferences have sprung up
- Field brought many people together
- Not perfect – but we are achieving success even with remarkably small budgets

1.3.4 *Seamus Ross*

- Depends on measures of success
- Really 25 years old from Hedstrom paper and Bearman paper in 1989
- Engaged researchers
- Models and frameworks there
- Recognize complexity of issues
- Created new knowledge and developed new policies
- Scalable, automated, ubiquitous solutions not there – integrated into system design
- Foggy and lacking in the solutions space
- Need more public imagination
- Digital preservation still a standalone/post-ingest activity
- Automation – much depends on ability to transform activities from niche research to large scale recognition

1.3.5 *Ross Wilkinson*

- Complexity problem can be broken down in objects with identification
- Persistent ID and locus for preserving things is important

- People who really need preservation to occur (senior people who have influence over dollars) are really caring
- Community of data archives who care is out there

1.3.6 Barbara Sierman

- Practical point of view - zipper
- After invention, it was more than 30 years before zipper became a commodity (buttons, hooks and eyes were around).
- Metaphor of zipper falling apart (removing zipper before washing, twice as expensive, rust, etc.)
- We don't have that time like the zipper
- Need large group of stakeholders, spending money on it.
- Three main issues
    - Unable to frame the message for a larger audience. Need better terminology (for management, but also for Europe for researchers, publishers, industry, etc.). Too much jargon.
    - Hiding our failures – we need a shift there
    - No proper toolkit - we should have that

1.4 Closing remarks

- Barbara – need more practical solutions (what can I do tomorrow? Except for collaboration)
- Seamus – wider recognition of the importance of preservation at all educational levels (e.g. grade schoolers on preservation of access to digital photos).
    - Automation of workflows and processes (similar to automotive industry)
    - More significance of appraisal and selection
    - Ensure digital preservation functionality is built into design and development – haven't made much progress
- Helen – people do collect and use data, but we haven't seen any recognition of stewardship functionality in these steps
- Ed – need better tools
    - Shared gap analysis and road mapping
    - More evidence on how much distributed software development costs
    - Sophisticated understanding with public/private partnership
    - Providing support for emerging skills – more easy to use tools, more internships, more demystifying problem
    - More cross-domain collaboration (e-government, industry, GLAM, data management, etc.)
- Andi
    - Show us one domain/discipline where if you get 200 people in the room, and ask if they are happy with the tools
    - There is no perfect, stable state. There never has been
    - Move beyond cultural heritage domain – we have borne the burden of taking us where we are now. Other disciplines are benefiting from it now. Reach out to other industries to take on burden. Like ERPANET reaching out to pharmaceuticals, etc.
    - We need to adopt a new language
    - Dare to think beyond standard topics
- Closing remarks from Shaun
    - Definition of value –
        - Patents as example – public/private benefit – public gives you the private right to that innovation for public benefit
    - Collaboration
        - Roadmapping could drive collaboration (what challenges will we be facing in two years we need to face)
        - But is it even possible to roadmap when the world is changing so dramatically?
    - What's the buzz word? – big data question from Janet Delve
        - Embrace the buzz words and use them to your own benefit
        - Using them to tell our stories to the public/business/industry

Closing Remarks

A Data-Geek’s Perspective on iPres 2014

Andrew Treloar
Australian National Data Service
Monash University Caulfield Campus
Caulfield East, 3145, Australia
+61 3 990 20572
andrew.treloar@ands.org.au

When I was asked by the programme committee to summarise the conference, my first reaction was to ask “are you sure?”. My second reaction was to accept with gratitude!

The reason for my initial reticence is that iPres is not my community. You should also be aware that I was only able to attend part of the conference (because of the multiple tracks), although having the proceedings on the lanyard was very useful for the sessions I couldn’t attend. Finally, because of my job, I tend to see things through a data lens and this probably coloured some of my reactions.

Let me begin by complimenting the programme committee on the quality of the papers and keynotes that they attracted, and all those who attended on the level of enthusiasm and interaction. It is clear that you all care passionately about preservation, and this showed.

I thought the conference had a great selection of practice papers (particularly in the short papers section). These showed practitioners reflecting thoughtfully and intentionally on what had worked and what hadn’t. I also applaud the number of speakers arguing for pragmatic solutions that don’t try to be perfect – this is a shift from preservation events I have attended in the past. There was also a recognition (in the data domain at least) that doing it perfectly (or even well?) is impossible – pragmatism is the only appropriate response.

On the subject of data, there were a number of talks (keynote and otherwise) about the importance of data to the scholarly record. These, either implicitly or explicitly, argued for the importance of preserving that data and the processes that produced it. This is, I think, a new frontier for many within the preservation community. There are a whole series of new challenges in the research data space – it is not the same as the existing born-digital challenge, for reasons explored in the paper that Herbert van de Sompel and I presented.

Informed by my experiences in the eresearch infrastructure domain over the last decade, I would encourage those people who are building tools to avoid the temptations of reinventing wheels where perfectly good ones exist already. There is real value in adding effort to an existing community of developers, and it results in more sustainable outcomes.

On the subject of sustainability, I would again commend the poster by Paul Wheatley on lessons learned in developing digital preservation tools.[1] He said everything I was planning on saying on the subject, and said it better:

- engage with the community
- build on existing work
- design for longevity
- ally with a custodian

I would also argue for a stronger focus on user-pull (and development based on well-defined and grounded use cases) over technology push. Having said that, I did seem some encouraging signs at the conference of a desire to build on what is there and meet the needs of real users, as well as some interesting research ideas that may bear fruit in the future.

Let me conclude by reminding of something that I am sure you all know: Digital preservation is too important not to care about it. Much of the work reported at this conference will play a key role in the solutions that need to be developed. Thank you for your commitment and energy in developing those solutions!

[1] http://www.slideshare.net/prwheatley/ipres2014-poster-02

Workshops and Tutorials

Defining a Roadmap for Economically Efficient Digital Curation – a 4C Project Workshop

Neil Grindley
JISC
Brettenham House (South Entrance)
5 Lancaster Place
London WC2E 7EN
n.grindley@jisc.ac.uk

Katarina Haage
Deutsche Nationalbibliothek
Information Technology
Adickesallee 1 D-60322
Frankfurt am Main
k.haage@dnb.de

Paul Stokes
JISC
One Castlepark Tower Hill
Bristol BS2 0JA
p.stokes@jisc.ac.uk

ABSTRACT

The 4C Project is tasked with delivering a Roadmap report and it is this drive towards 'economic efficiency' in relation to digital curation that will be central to the agenda that it sets out. This workshop is an important opportunity to connect with stakeholders and get input for a critical deliverable of the project. But it is also an opportunity for participants to learn more about the economics of digital curation and to critically assess the efficiency and sustainability of their own services and solutions.

General Terms

Communities, strategic environment, digital preservation marketplace, theory of digital preservation.

Keywords

Economics, policy, strategy.

1. INTRODUCTION

The 4C Project (a Collaboration to Clarify the Costs of Curation) is a European Commission funded two year coordination action which has been funded to provide useful, useable resources that provide better support to identify and quantify the cost of digital curation. From the outset, however, the project has taken the view that costs cannot be dealt with in isolation from a number of other related concepts (e.g. benefits, risk, quality, sustainability) and this holistic view might more accurately be described as an economic perspective on digital curation.

Borrowing the language of economics and mapping it onto digital curation needs to be done selectively and carefully. Digital assets do not have the same attributes as other kinds of (financial) assets and equally, it may not be possible to define when digital assets become (economic) liabilities in any objectively quantifiable way. However, there is still terminology from the field of economics that may help to define what the digital curation community might aspire to over the next few years and the starting point for this workshop is the concept of 'economic efficiency'- which might be defined as the optimised situation where it is no longer possible to add quantity or value given a finite availability of resources.

The 4C Project is tasked with delivering a Roadmap report and it is this drive towards 'economic efficiency' in relation to digital curation that will be central to the agenda that it sets out. The consultation, stakeholder engagement, analysis and modelling work that have been done allow some principles to be proposed and some assertions to be made that will form the backbone of the report. The purpose of a Roadmap – particularly where it seeks to set out an action agenda for a range of stakeholders across various communities – is to make politically astute observations and to arrive at plausible conclusions. This is only possible via early interaction with stakeholders and by achieving some level of community validation before publication and this is the purpose of the workshop. One of the guiding principles of the 4C Project is to create a better understanding of the economics of digital curation through collaboration; and also to be an 'open and social' project and to listen to the needs of the community. iPRES 2014 occurs at roughly the three quarter point of the two year project and provides a timely opportunity to check and refine the draft Roadmap.

Early ideas and discussions about the structure and content of the Roadmap have indicated that it will need to address various questions.

- What vision should we advocate and what principles should we espouse to bring about economically efficient digital curation?
- What current economic inefficiencies do we need to eliminate?
- What or who is the most influential mechanism to bring that about and where will that influence most be felt?
- What is the policy, business and regulatory framework for digital curation and how is it likely to change?
- Over what timescales should we advocate action?
- How can we most economically sustain and exploit existing work? (including the 4C Project outputs)
- How are the economic requirements of stakeholders changing?
- Is it possible and economically desirable to try and align digital curation practice (including standards and terminology)?
- How can we most effectively invest in digital curation at the institutional, national and international level?

This workshop is an important opportunity to connect with stakeholders and get input for a critical deliverable of the project. But it is also an opportunity for participants to learn more about the economics of digital curation and to critically assess the efficiency and sustainability of their own services and solutions.

2. Intended content

Half day workshop

09.00 – Introduction to the aims of the session, the purpose of the 4C Roadmap and a perspective on the economics of digital curation (presentation and Q&A)

09.30 – Presentation of the 4C project outputs

10.00 – Breakout groups to discuss (and then briefly feedback on) digital curation economic needs & gaps

10.30 – Break

10.45 – Presentation of the draft Roadmap for Economically Efficient Digital Curation

11.15 – Breakout groups to discuss the Roadmap

12.00 – Feedback from the groups

12.45 – Summing up

13.00 – Lunch

Open to public

This workshop will be paid for by the 4C Project and be open to all participants interested in the economics of digital curation.

Requirements for its organisation

All that is required is a room, a screen and a data projector. The ideal audience would be a mixture of those with opinions and information to offer the 4C project and those who would take information back to their organisation and prepare the way for effective dissemination towards the end of the project. This is both an input and an output opportunity. A workshop of anywhere between 10-20 people would be a useful size and a great opportunity to have a detailed conversation with an interested audience.

Local capability

See above

Speakers

The lead for the workshop will be the 4C Project coordinators Neil Grindley and Paul Stokes. Other 4C project partners will contribute and if possible, affiliate stakeholder organisations will also present.

Intended Audience

Practitioners, Managers and Funders – this has applicability at all levels and should be of practical, tactical and strategic interest.

3. WORKSHOP OUTCOMES

This workshop was the first opportunity to get face-to-face feedback from the community on the draft 4C project roadmap. ‘Investing in Curation: a shared path to sustainability’ states six messages and sets out a number of actions that various stakeholder groups should act upon to realise a suggested shared vision that could be realised by the year 2020.

The draft Roadmap is available at: http://4cproject.eu/d5-1-draft-roadmap

The vision is as follows:

> *In five years time (2020) it will be easier to design or procure more cost effective and efficient digital curation services because the costs, benefits and the business cases for doing so will be more widely understood across the curation lifecycle and by all relevant stakeholders. Cost modelling will be part of the curation planning and management activities of all digital repositories.*

The workshop was divided into two main sections. Firstly participants were asked to consider the main challenges they and their institutions faced with curating digital assets (particularly in relation to economic issues). Secondly, they were asked to think about the draft 4C Roadmap messages and to consider how relevant they were to their own local context and to what extent they were plausible and sensible as an agenda for action and change.

The first discussion (challenges) surfaced the following issues:

- The scale and type of issues that will need to be faced is difficult to predict but international collaboration and knowledge exchange will mitigate the impact of that uncertainty
- There are important stakeholders (e.g. certain areas of government and publishing) who don’t yet feel that curation planning is their problem or who don’t yet understand that ‘digital is not technology’. Or to put it another way, they haven’t yet understood that digital assets are a business issue and not an IT problem.
- We need better models to understand the cost of collaboration; and to understand the scale and costs of the R&D that may be needed
- There are ownership issues that cause problems around the openness (or not) of data; about how to define the costs of distributed costs centres; who actually owns digital collections; and monolithic IT budgets that can’t be broken down into departmental figures.
- Human & managerial issues (rather than technical) require additional focus and resource
- Joined up infrastructure is expensive but is a requirement
- Sustainability is a big challenge and this has to be tackled by robust business and use cases; through automation rather than manual curation processes; and by making the activities (and the assets) more visible and apparent to the organisation
- Selection is happening but techniques need to evolve to cope with appraisal at scale
- The current software solutions are inadequate so demand and requirements need to be better articulated and tools need to be more carefully specified
- There is a lot of inertia and inflexible legacy working practices within organisations that slow down ingest; limit file format choices; hinder policy development and changes to working practices
- Finding properly qualified staff and the right kind of curation expertise is hard

The second discussion prompted the following thoughts in response to the Roadmap:

- The focus of the Roadmap is very much on the ‘asset’ nested within an ‘organisation’; structures may change over time and an alternative or additional focus might be on people and skills and emerging technologies

- Many organisations (especially libraries and archives) are still very wrapped up in dealing with analogue collections and the transition to digital and the curation challenges associated with this
- Predicting 5 years into the future is a long or a short time depending on organisational context; the predictions for 2025 in the Roadmap are already being tackled in practice now
- Message 1 ('Make choices and select') was one of the more problematic statements. Selection may be incompatible with 'big data' techniques and may also be in conflict with the mission of some libraries; but it may also be stating the obvious or rehearsing accepted practice in environments where digital curation is established
- Message 2 ('Demand efficient systems') skews activity towards procurement rather than in-house development and assumes that there is already an effective marketplace and market analysis that can be drawn upon
- Message 3 ('Build scalable infrastructure') was an uncontroversial message
- Message 4 ('Sustainability') should extend beyond thinking about organisations and assets and should also include software and applications and embedding sustainability into up-front funding arrangements
- Message 5 ('Make funding dependent on lifecycle costing') should be clearer about what the funding will actually support and be wary of inhibiting activity entirely
- Message 6 ('Be transparent and share') should reference the power of open source and other 'open' concepts and emphasise the potential to improve quality
- There are general issues with definitions throughout the Roadmap, for example it may not be clear to everyone what is meant by, 'lifecycle', 'value' and 'efficient' in the context they are used
- There are important contextual organisational differences that need to be acknowledged, particularly in cases where assets are generated internally or acquired from external sources; and where activity is community-led or where it is commercially-driven
- Curation and preservation thinking needs to happen at the content (assets) level but also at the application (systems) level and at the platform (environment) level and this has economic implications
- The issue of standards alignment and the convergence of practice is complicated and it is not clear whether it is an opportunity or a problem and how the economics work out in terms of community practice and functional markets
- The roadmap needs to be clear about the ownership problem (see 'challenges' above) and who should be taking responsibility and in what context
- There is much that can be learnt and taken from business and big data industries; public sector organisations should be more open to these ideas to introduce more economic practices
- The Roadmap could set out more of a research agenda and provide an innovation platform for students and early-career researchers

Born Digital Appraisal, Ingest, and Processing

Jessica Moran
National Library of New Zealand
PO Box 1467
Wellington 6140 New Zealand
64 4 460 2862
jessica.moran@dia.govt.nz

Leigh Rosin
National Library of New Zealand
PO Box 1467
Wellington 6140 New Zealand
64 4 474 3109
leigh.rosin@dia.govt.nz

ABSTRACT

This workshop is designed to gather together the community of digital archivists and others working specifically with born digital collections to discuss aspects of the appraisal, processing, and ingest process.

General Terms

Preservation strategies and workflows, specialist content types, case studies and best practice, and training and education.

Keywords

Born digital, digital archivists, appraisal, ingest workflows.

1. INTRODUCTION

The purpose of the workshop is to gather together a community of digital archivists, and others working specifically with born digital collections to discuss the appraisal, processing, and ingest process, to share tools and workflows, and to begin to discuss the articulation of best practices.

2. WORKSHOP OVERVIEW

The workshop will bring together experienced digital preservation practitioners, specifically digital archivists, digital curators, and other practitioners with a responsibility for, or interest in, born digital preservation workflows and systems. A number of digital archivists will give short presentations sharing their experiences and insights around particular aspects of working with born digital collections, including initial technical appraisal, media transfer, ingest workflows, and specific tools. Discussion will focus around these workflows and tools, as well as issues or problems encountered, successes and failures, lessons learned, and future planning.

The workshop is designed to further the articulation of international current and best practices for digital archivists. The workshop will deal explicitly with the technical appraisal and ingest of born digital materials. As the field of born digital collecting and preservation has grown, the body of knowledge and experience in this area has also grown and there is a greater need than ever to share that knowledge among a wider group of practitioners.

The workshop will bring together a number of practitioners and presenters and will include:

- Leigh Rosin, Digital Archivist, National Library of New Zealand
- Douglas Elford, Emma Jolly, and Somaya Langley, Digital Collecting, Australian National Library
- Donald Mennerich, Digital Archivist, New York University
- Cal Lee Associate Professor, School of Information and Library Science, University of North Carolina, Chapel Hill,
- Erin O’Meara, Gates Archive.

The workshop will combine formal presentations, contributions from participants, and group discussion. As such, the workshop would best benefit those who already have experience as practitioners in the area of digital preservation workflows and systems. It is hoped the workshop will contribute to further networking and information sharing among digital archivists, curators, and others with a responsibility or interest in born digital preservation.

3. WORKSHOP OUTCOMES

The workshop was attended by 48 participants. Participants were lead through a brainstorming session in small groups. Each group generated one key issue or problem they would like to see examined in more detail, relating to digital ingest and processing. These were:

- Difficulties relating to capability-building and skills retention in digital teams. Lack of succession-planning
- Multi-disciplinary communication and creating networks. How to effectively share resources and experiences.
- Archivists who code and how to bridge the gap between archives/libraries and computer science
- Donor relationships and training. Building effective techniques for dealing with donors of born digital collections
- Metadata interoperability. We have metadata, now how to we effectively link and use it
- We need a workflow management tool to help us track all these ingest activities performed by various staff. Do we look to ticketing systems? Project management tools? Acquisitions systems?

Workshop participants engaged in a discussion about these issues. No definitive “answers” were provided, but participants were able to explore the issues in depth and gain a better understanding of the problem space.

PREMIS Implementation Fair Workshop

Peter McKinney
National Library of New Zealand
Cnr Molesworth & Aitken St
Wellington
New Zealand
peter.mckinney@dia.govt.nz

Eld Zierau
The Royal Library of Denmark
Søren Kierkegaards Plads 1
DK-1016 København K
+45 91324690
elzi@kb.dk

Rebecca Guenther
Library of Congress/Consultant
101 Independence Ave SE
Washington, DC 20540 USA
+1 703 298 0157
rguenther52@gmail.com

ABSTRACT

This workshop provides an overview of the PREMIS Data Dictionary for Preservation Metadata, a standard addressing the information you need to know to preserve digital content in a repository. It includes an introduction to PREMIS and reports from the preservation community on implementation of the standard in various systems or contexts.

General Terms

infrastructure, preservation strategies and workflows, case studies and best practice, preservation strategies and workflows.

Keywords

Preservation metadata, Preservation repository implementation, Data dictionary

1. INTRODUCTION

The PREMIS Implementation Fair Workshop is one of a series of events organized by the PREMIS Editorial Committee [1] and that has been held in conjunction with previous iPRES conferences.

At iPRES 2014, the workshop will give the audience a chance to understand the PREMIS data dictionary and give implementers, and potential implementers, of the *PREMIS Data Dictionary for Preservation Metadata* an opportunity to discuss topics of common interest and find out about latest developments.

2. OUTLINE OF WORKSHOP CONTENT

2.1 Overview of the PREMIS Data Dictionary

The *PREMIS Data Dictionary for Preservation Metadata* [2] is the international standard for metadata to support the preservation of digital objects and ensure their long-term usability. Developed by an international team of experts, PREMIS is implemented in digital preservation projects around the world, and support for PREMIS is incorporated into a number of commercial and open-source digital preservation tools and systems. This session provides an overview of the PREMIS Data Model (which was recently revised) and of the types of information specified to support the digital preservation process. Included will be a summary of the changes in version 3.0, which includes enhanced ability to describe intellectual objects and technical environments within the PREMIS context.

2.2 PREMIS Conformance

This session describes the work of the PREMIS Conformance Working Group and its effort to clarify what it means to adequately capture the essential metadata needed to support the essential functions of a digital repository. The group is pursuing two avenues of inquiry. The first has drafted conformance levels and is exploring what metadata is required for minimum conformance to PREMIS. The second explores the relationship between preservation metadata and functionality of a preservation system. These two avenues will together allow institutions to not only be able to understand their own conformance to PREMIS, but additionally reflect on how they utilise their metadata to drive, support and record preservation functions.

2.3 Implementation reports

Implementation reports will be solicited from the PREMIS Implementers community. Included will be a report on the National Library of Australia's implementation of a Tessella solution and one on the complexities of applying PREMIS to born digital data acquired on removable media.

3. WORKSHOP SERIES

The PREMIS Implementation Fair at iPres 2014 will be the sixth in a series and have been held in conjunction with iPres since 2009. These events are intended to highlight PREMIS activities, discuss issues concerning implementation, and provide a forum for implementers to discuss their activities, issues, and solutions. Because this is a rapidly changing area, it is important to provide continuous updates.

4. INTENDED AUDIENCE

The workshop is designed for those involved in selecting, designing, or planning a preservation project or repository using preservation metadata. This includes digital preservation practitioners (digital librarians and archivists, digital curators, repository managers and those with a responsibility for or an interest in preservation workflows and systems) and experts of digital preservation metadata and preservation risk assessment.

5. SHORT BIOGRAPHIES OF ORGANIZERS

Peter McKinney is the Policy Analyst for the Preservation, Research and Consultancy programme at the National Library of New Zealand Te Puna Mātauranga o Aotearoa. He is a member of the PREMIS Editorial Committee and part of the Conformance Sub-Committee. Most recently he has been coordinating work on the NSLA Digital Preservation Technical Registry.

Eld Zierau has been a member of the PREMIS Editorial Committee since 2013. She is a digital preservation researcher and specialist, with a PhD in digital preservation. She is a computer scientist, and has worked with almost all aspects of IT in private industries for 18 years, before starting in digital preservation in 2007. She has been working with many aspects of digital preservation, and she is involved as an architect or a consultant on major initiatives such as a new digital repository including data modeling of metadata for preservation.

Rebecca Guenther is Chair of the PREMIS Editorial Committee, on which she has served since its establishment in 2006. She worked at the Library of Congress on metadata standards in the Network Development Office for 22 years and is currently an independent consultant in New York on metadata development and training; she also continues to work part-time for the Library of Congress. She was co-chair of the original PREMIS Working Group which developed the *PREMIS Data Dictionary for Preservation Metadata.*

6. PROCESS FOR SOLICITING CONTRIBUTIONS

Contributions will be solicited from the PREMIS Implementers' Group via its discussion list (pig@loc.gov). To subscribe go to: http://listserv.loc.gov/listarch/pig.html. The PREMIS Editorial Committee will review all requests. After workshop proposal is approved, a call will be sent for contributions to the implementation portion and deadline will be within a month.

7. WORKSHOP OUTCOMES

7.1 Participant expectations:

Participants were invited to introduce themselves and their motivation for joining the workshop. Most participants described no working knowledge of PREMIS but wanting to learn more in order to

1) be able to implement the standard in their institution
2) gain better personal understanding of requirements for preservation metadata
3) be able to better explain the needs for PREMIS within their institution

7.2 Questions

Q: Can I dump technical metadata elsewhere in the system and just include a pointer towards that place in PREMIS?
A: Yes.

Q: Are events preservation events only or really any events related to the object?
A: Per definition events can be related to creation, modification and access. Most events related to the object should be able to be mapped to those groups. The institution needs to define whether an event is indeed a preservation event to them or not.

Q: We run fixity once a month, should we keep the info?
A: It is recommended to capture information about any events that touch an object, but it is up to the institution to define how this is realized. It would be possible, for example, to only capture the info if something goes wrong.

Q: Can we capture even more information in PREMIS, like descriptive metadata?
A: In general extensive descriptive metadata should be captured elsewhere. The specific requirements of the collection should be looked at to decide what to capture where. It would be helpful for the PREMIS committee to have some concrete examples from the user community.

Q: Are there any recommendation and guidance as to what to include in the extensions to ensure long-term understandability by e.g. including fixed vocabulary or standards?
A: No recommendations and guidance are available. It is implied that standards should be used for the extensions – however, this is of course hard in cases such as the eventOutcomeDetails. It would be helpful for the PREMIS committee to have some concrete examples from the user community.

Q: Where can I dump my extensive ffmpeg output? Does it go into eventOutcome or elsewhere?
A: It is important to differentiate between tool output and problem reporting by tools. EventOutcome should capture if the event ran ok or not and include error messages.

Q: How do large institutions like national libraries deal with schema changes – e.g. in the case of the upcoming PREMIS v3?
A: At first check how you can implement this in your system. If you have a vendor, talk to them early on. It is also important to think about what to do with the preserved objects – to change the preservation metadata for those, you could consider a tool-based approach or running them through the entire system again.

Q: Why don't you directly describe the policies in PREMIS (in relation to Eld's presentation on preservation level)?
A: Because the policy may change regularly.

Q: Has PREMIS looked at incorporating the SCAPE controlled vocabulary for policies?
A: Not that we know. But will be recommended to the Committee.

Q: Has the environment extension been tested? Will it ensure preservation and renderability? Who is preserving all these environments?
A: That's a general digital preservation questions – it is good if we can point towards registries for a lot of this.

7.3 Implementation examples:

Eld Zierau presented the PREMIS implementation at the Royal National Library of Denmark.

Scott Wajon (State Library of New South Wales) brought in an example of a metadata file the institution received from a service provider. The file included PREMIS and MIX metadata. It was used to look at what kind of information could be captured from external processes. The file was interesting in that only event metadata was codified in PREMIS semantic units (why had the vendor made that decision). In particular, the file included extensive information about a deskewing event. It was discussed how this information could be relevant depending on whether it was preformed on a master or on a derivative file. Explicit information about the software/agent which was used to perform the event should be included.

Michelle Lindlar presented work being done as part of the DURAARK project in a pre-ingest workbench for architectural 3D data. Regarding a PREMIS implementation in the workbench process, three questions were formulated:

1. If the pre-ingest workbench runs externally (e.g. as a service) with no knowledge of the preservation repository, is it still an agent or is it something else?
2. As the pre-ingest workbench is a complex system combining multiple tasks and wrapping separate tools for e.g. file format identification and metadata extraction, is it a series of agents or something else?
3. Within DURAARK, "a building / structure" is considered an intellectual entity. Representations of the entity always stand in temporal / spatial relationships and dependencies – i.e., scans from different years or plans describing pre-/post-refurbishing states. These representations should therefore be rather positioned at an IE level, calling for a nested IE structure. Is this possible within PREMIS and are there known reference implementations for this?

It was discussed how a pre-ingest workbench can be described using the environment entity in PREMIS v3. The DURAARK workbench can therefore be seen as a nice use case for this new entity, where an external system is described which produces a SIP (and therefore generates a PREMIS file) to be deposited into an institution's digital preservation system. The environment entity also allows for the detailed description of the different agents involved within the pre-ingest process.

The nested structure is possible in theory, however, no reference implementation is known.

7.4 Action Items

- Put all slides from the event onto PREMIS website.
- Put on website (and send to participants) sample METS showing PREMIS used for newspaper digitisation work.
- Investigate SCAPE controlled vocabularies.
- Editorial Committee to investiage enriching advice on container extensions (in particular eventOutcomeDetail).

8. REFERENCES

[1] PREMIS Maintenance Activity, http://www.loc.gov/standards/premis/

[2] PREMIS Editorial Committee. 2012. *PREMIS Data Dictionary for Preservation Metadata* (Library of Congress). http://www.loc.gov/standards/premis/v2/premis-2-2.pdf

ICA-AtoM, Archivematica and Digital Preservation

Dr Lise Summers
State Records Office of Western Australia
James St,
Perth WA 6000,
Australia
Lise.summers@sro.wa.gov.au

Meg Travers
State Records Office of Western Australia
James St,
Perth WA 6000,
Australia
meg.travers@sro.wa.gov.au

ABSTRACT

In this paper, we describe the formatting guidelines for iPRES 2014 Proceedings.

General Terms

communities, preservation strategies and workflows, specialist content types.

Keywords

Government business systems

1. INTRODUCTION

ICA-AtoM is an open source, web-based archival description software application that is based on International Council on Archives standards. The first version of it was released in 2008 with funding from a variety of organisations including UNESCO, World Bank Group Archives, the Dutch Archiefschool, the Direction des Archives de France and the United Arab Emirates Centre for Documentation and Research. In 2013, the State Records Office of Western Australia (SROWA) has invested in further development of the software, chiefly to include support for the Australian Series Registration System as well as simple preservation workflow. These additions to ICA-AtoM will be completed in the middle of 2014 and made freely available as part of the open source package to any archives wishing to download the software.

This work has aligned the Australian Series Registration System within an international standard ISAD compliant system, and will go some way to eventually bringing the two together.

This workshop will provide an overview of using ICA AtoM with special attention to archival description using the Australian Series Registration System; ingest into the complementary Archivematica digital preservation system and attaching digital objects to AtoM.

The workshop is open to the public.

2. ORGANISERS

The State Records Office of Western Australia is the Western Australian public records authority with responsibility for managing, preserving and providing access to the State's archive collection.

3. OUTLINE

Length: 3 hours

Outline of Content

Time	Subject	Presenter
30 mins	Introduction: Why AtoM? Why open source?	Meg Travers
60 mins	Archival description in AtoM	Lise Summers
15 mins	*Break*	
30 mins	Digital objects – access vs preservation – uses and workflows	Meg Travers
40 mins	Digital preservation and digital access in AtoM and Archivematica	Lise Summers and Meg Travers

4. PRESENTERS AND PANELLISTS

Dr Lise Summers (@morethangrass) is employed at the State Records Office of Western Australia as a Senior Archivist, where she has been the Preservation Program co-ordinator since 1998, and is closely involved in their ICA-AtoM project. Lise also lectures at Curtin University in the School of Information Studies, teaching in the areas of archives management and conservation and preservation. Lise is active in the fields of history and heritage, being the current President of the History Council of WA. Her PhD thesis, 'From wasteland to parkland: a history of designed public open space in the City of Perth, 1829 – 1965', was awarded by the University of Melbourne in 2008.

Meg Travers (@museit) is the Manager Digital Archives at the State Records Office of Western Australia and has project managed the work undertaken by SROWA on ICA-AtoM. She has worked in information technology for over 20 years, the last 10 in the GLAM sector. Meg is a post graduate student at the Western Australian Academy of Performing Arts researching the preservation of early electronic musical instruments, and is currently recreating a Trautonium, one of the very first electronic instruments invented. She is also an active composer and performer of electronic music, and is the go-to person in WA for questions on archaic music technology.

Preserving Data to Preserving Research: Curation of Process and Context

Angela Dappert
Digital Preservation Coalition Senior Project Officer
DPC c/o The British Library,
Floor 5, Room 14, 96 Euston Road,
London NW1 2DB +44 (0) 20 7412 7028
angela@dpconline.org

Rudolf Mayer, Stefan Pröll, Andreas Rauber
Secure Business Austria,
Vienna, Austria
mayer@ifs.tuwien.ac.at
sproell@sba-research.org
rauber@ifs.tuwien.ac.at

Raul Palma
Poznan Supercomputing and Networking Center,
Poland
rpalma@man.poznan.pl

Kevin Page
University of Oxford e-Research Centre,
United Kingdom
kevin.page@oerc.ox.ac.uk

Daniel Garijo
Universidad Politecnica de Madrid,
Spain
dgarijo@delicias.dia.fi.upm.es

ABSTRACT

Awareness of the need to provide digital preservation solutions is spreading from the core memory institutions to other domains, including government, industry, SME and consumers. In many of these settings we are, however, faced with preserving more than just data. In the domain of eScience, for example, investigations are increasingly collaborative. Most scientific and engineering domains benefit from building on the outputs of other research by sharing information to reason over and data to incorporate in the modeling task at hand.

This raises the need for preserving and sharing entire eScience workflows and processes for later reuse. We need to define which information is to be collected, create means to preserve it and approaches to enable and validate the re-execution of a preserved process. This includes and goes beyond preserving the data used in the experiments, as the process underlying its creation and use is essential.

The TIMBUS project and Wf4Ever project team up for this half-day tutorial to provide an introduction to the problem domain and discuss solutions for the curation of eScience processes.

General Terms

Infrastructure, preservation strategies and workflows

Keywords

e-Science, data preservation, workflows, semantics, Research Objects, Context Models

1. TUTORIAL STRUCTURE

The tutorial will cover the following topics:

Introduction to Process and Context Preservation: The introduction will motivate the need for process and context preservation, illustrate how this task is difficult in an evolving domain, and introduce a common use case, based around the work of a researcher in Music Information Retrieval [1], which is used in the rest of the tutorial to illustrate approaches and tools for the rest of the tutorial to illustrate approaches and tools.

Data Citation: Data forms the basis of the results of many research publications, and thus needs to be referenced with the same accuracy as bibliographic data. Only if data can be identified with high precision can it be reused, validated, verified and reproduced. Citing a specific data set is not trivial, however: it exists in a vast plurality of specifications and instances, can potentially be huge in size, and its location might change. We will provide an overview over existing approaches to overcoming these challenges. We will also present the issue of creating data citations of data held in databases, especially of dynamic data sets where data is added or updated on a regular basis.

Re-usability and traceability of workflows and processes: The processes for creating and interpreting data are complex objects. Curating and preserving them requires special effort, as they are dynamic, and highly dependent on software, configuration, hardware, and other aspects. We will discuss these issues in detail, and provide an introduction to two complementary approaches.

The first approach is based on the concept of Research Objects, which adopts a workflow-centric approach and thereby aims at facilitating the reuse and reproducibility. It allows us to package the data, along with the scientific context information of how these resources were used or produced, as one Research Object, and thus to share and cite it. This enables publishers to grant access to the actual data and methods that contribute to the findings reported in scholarly articles.

A second approach focuses on describing and preserving a process and the context it is embedded in. The artifacts that may need to be captured range from data, software and accompanying documentation, to legal and human resource aspects. Some of this information can be automatically extracted from an existing process, and tools for this will be presented. Ways to archive the process and to perform preservation actions on the process environment, such as recreating a controlled execution environment or migration of software components, are presented. Finally, the challenge of evaluating the re-execution of a preserved process is discussed, addressing means of establishing its authenticity.

2. INTENDED AUDIENCE

The tutorial is targeted at researchers, publishers and curators in eScience disciplines who want to learn about methods of ensuring the long-term availability of experiments forming the basis of scientific research.

3. EXPECTED LEARNING OUTCOMES

The tutorial participants will become familiar with:

- Motivations and challenges of process preservation;
- Motivations, stakeholders and challenges of making data citable;
- How data is cited today, best practices, guidelines and metadata standards;
- Available technologies for identifiers: Archival Resource Key (ARK), Digital Object Identifiers (DOI), Extensible Resource Identifier (XRI), HANDLE, Life Science ID (LSID), Object Identifiers (OID), Persistent Uniform Resource Locators (PURL), URI/URN/URL, Universally Unique Identifier (UUID);
- Approaches and Initiatives for citing data: CODATA, Data Cite, OpenAire, challenges and opportunities: granularity, scalability, complexity and evolving data sets current research questions;
- Ontologies needed to capture research objects: Core Ontology of the RO family of vocabularies, workflow centric ROs, provenance traces, life cycle of research objects;
- Wf4Ever Toolkit / technological infrastructure for the preservation and efficient retrieval and reuse of scientific workflows: software architecture, functionalities, software interfaces to functionalities, reference implementation as services and clients:
  - Collect, manage and preserve aggregations of scientific workflows and related objects and annotations
  - Workflow sharing through a social website
  - Execution of workflows;
  - Testing completeness, execution, repeatability and other desired quality features;
  - Testing the ability of a Research Object to achieve its original purpose after changes to its resources;
  - Recommendations of relevant users, Research Objects and their aggregated resources;
  - Converting workflows into Research Objects;
  - Search for workflows by input parameters or frequency of use;
  - Collaborative environment;
  - Access and use of research objects and aggregated resources;
  - Synchronization with remote repositories;
  - Visualization of research object evolution;
- TIMBUS context model and tools to semi-automatically capture the relevant context of a business process for preservation:
  - The scope of context regarding business process preservation - technology, application and business context, aligned with enterprise architecture;
  - The context meta-model, with domain independent and domain specific aspects;
  - Demonstration of a context model instance of example processes (in the eScience domain);
  - Tools to automatically capture some parts of the context (software dependencies, data formats, licenses, etc);
  - Outlook on reasoning and preservation planning, based on the context model.

4. BIOGRAPHY OF THE PRESENTERS

Angela Dappert is Head of Research and Practice at the Digital Preservation Coalition. She also serves on the PREMIS Editorial Committee. In both capacities she is involved with the issues of modelling and defining metadata for computer environments. She has worked at the British Library on data carrier stabilization, digital asset registration, digital preservation planning and characterization, eJournal ingest, and digital metadata standards. Before this she worked for Schlumberger, the University of California, Stanford University and Siemens. Angela holds a Ph.D. in Digital Preservation from the University of Portsmouth, an M.Sc. in Medical Informatics from the University of Heidelberg and an M.Sc. in Computer Sciences from the University of Texas at Austin.

Daniel Garijo is a PhD student in the Ontology Engineering Group at the Universidad Politecnica de Madrid. His research activities focus on e-Science and the Semantic Web, specifically on how to increase the understandability of scientific workflows using provenance and metadata. He has been a member of the W3C Provenance Working Group, and previously participated in the Wf4Ever project.

Rudolf Mayer is a researcher at Secure Business Austria, as well as the Department of Software Technology and Interactive Systems at the Vienna University of Technology. His research interests cover digital preservation, specifically the preservation of processes, information retrieval (specifically on text documents and music), data analysis and machine learning. He has many years of lecturing experience in these subjects. He has been involved in the DELOS and PLANETS projects, and currently works on digital preservation aspects in the FP7 projects APARSEN and TIMBUS.

Kevin Page is a researcher in the Oxford e-Research Centre, University of Oxford, UK. His work on web architecture and the semantic annotation and distribution of data has, through participation in several UK, EU and international projects, been applied across a wide variety of domains including sensor networks, music information retrieval, clinical healthcare, and remote collaboration for space exploration. His current research focuses on the application of semantic web architecture to information management systems for scientific workflows, musicology, and social machines, and the common approaches that underly these seemingly disparate subjects. He has previously organized and presented tutorials at the Extended Semantic Web Conference, the International Society for Music Information Retrieval conference and the Oxford Digital Humanities Summer School.

Raul Palma is a researcher at Poznan Supercomputing and Networking Center (PSNC). His research interests cover digital preservation, particularly of scientific methods, provenance and evolution of digital artifacts, ontology engineering and distributed technologies. He has participated in several EU projects, including the Network of Excellence Knowledge Web, NeOn, e-Lico and WF4Ever. He has many years of lecturing experience in related topics, both at the university and private institutions. He has authored or co-authored several vocabularies and ontologies, such as the Research Object evolution Ontology, Ontology Metadata Vocabulary (OMV) and different extensions for describing ontologies and related resources, models for collaborative ontology construction and digital multimedia repositories

Stefan Pröll is a researcher at SBA Research. His primary research focus lies on digital preservation, especially on security aspects of digital archives, including authenticity and provenance of digital objects. Further areas of interest are databases and data citation. Currently he is working on FP7 projects APARSEN and TIMBUS focusing on security and provenance related topics. Before he joined SBA in April 2011, he was working in international organizations in the area of Web development, Linux server and database administration.

Andreas Rauber is Associate Professor at the Department of Software Technology and Interactive Systems at the Vienna University of Technology. He is 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 has been involved in numerous initiatives in the area of digital preservation (DELOS, DPE, Planets, SCAPE, TIMBUS, APARSEN). He has been lecturing extensively on this subject at different universities, as part of the DELOS and nestor summer schools on digital preservation, as well as during a range of training events on digital preservation.

5. REFERENCES AND CITATIONS

[1] Kevin Page, Raúl Palma, Piotr Holubowicz, Graham Klyne, Stian Soiland-Reyes, Daniel Garijo, Khalid Belhajjame, and Rudolf Mayer. "Research Objects for Audio Processing: Capturing Semantics for Reproducibility." In Audio Engineering Society Conference: 53rd International Conference: Semantic Audio. Audio Engineering Society, 2014.

Memento
Uniform and Robust Access to Resource Versions

Herbert Van de Sompel
Los Alamos National Laboratory
PO Box 1663
Los Alamos, NM, USA
herbertv@lanl.gov

ABSTRACT

The Memento protocol tightly integrates the Web of the Present and that of the Past, making it possible to seamlessly navigate between both. The protocol defines an interoperable approach to access versions of a resource in web archives or content management systems such as wikis that leverage the URI of that resource and the datetime of the required resource version. Technically, the Memento protocol is an extension of HTTP that is fully based on the primitives of Web interoperability: URIs, resource representations, links, content negotiation. The tutorial will give an in-depth insight in various aspects of the Memento protocol that meanwhile has been published as RFC 7089.

General Terms

Infrastructure, preservation strategies and workflows, studies and best practice

Keywords

Versioning, web archives, content management systems, HTTP, content negotiation, interoperability, web persistence, internet robustness

1. TUTORIAL OUTLINE

The tutorial will provide a detailed insight in various aspects of the Memento "Time Travel for the Web" protocol. The tutorial is aimed to be useful for developers interested in implementing Memento compliant clients or servers, and project managers, information architects, repository administrators interested in learning whether and how Memento concepts can be used to meet challenges they face in the realm of resource versioning.

The remainder of this section details the focus areas of the tutorial.

1.1 Motivation

The tutorial will start by providing an insight in the motivation for the multi-year Memento effort, which is to be found in the poor integration between the Present and the Past Web. This lack of integration is exemplified by problems related to navigating from the current version of a resource to past versions, from past versions to the current version, and to consistently navigate the Web of the Past.

1.2 Memento Protocol

The effort to specify the Memento protocol started in late 2009 and concluded in December 2013 with the publication of the specification as an IETF RFC [1]. The core ingredients of the protocol will be introduced (datetime negotiation, Original Resource, TimeGate, TimeMap, Memento, Memento HTTP Headers) and the client-server interactions will be detailed for various patterns that differ mainly in whether an Original Resource, its TimeGate, and its Mementos reside on the same server or not. Special attention will be given to aspects of Memento Aggregation, which allows locating the temporally most appropriate archived resource version across web archives.

1.3 Memento and Resource Versioning

The Memento protocol is closely aligned with a common resource versioning pattern that consists of:

- Having a generic URI where at any moment in time the current version of the resource is accessible.
- Having a dedicated version URI for each resource version.

Systems that support this resource versioning pattern do not necessarily need to implement the entire protocol at once but can gradually implement aspects of it in a modular manner, with each step along the path providing increased functionality regarding access to resource versions. The incremental steps, as described in [2], will be explained:

- Providing HTTP response headers for resource versions to convey version date and links
- Publishing a TimeMap, a list of resource versions
- Exposing a TimeGate that supports datetime negotiation to access resource versions

1.4 Memento and Web Persistence

Memento's time travel capability provides an essential ingredient to address the well-known link rot, also known as "404 Not Found", problem. If a link is broken, follow it back into the past and obtain a version from a web archive or resource versioning system. But, as described in [3], in order to fully tackle the problem, several open questions remain to be answered, including: Which date should be used for time travel; How to convey information about a known archival version of a linked resource in an HTML page, and how to make sure such archival versions are created in the first place? The tutorial will provide insights in the thinking of two ongoing activities with this regard:

- Hiberlink[1], a Mellon-funded collaboration between the Los Alamos National Laboratory and the University of

[1] Hiberlink, http://hiberlink.org

Edinburgh that investigates the extent and nature of reference rot in web-based scholarly communication and explores approaches to ameliorate the problem. The project is inspired by a 2011 pilot study [4] that quantified scholarly link rot at an unprecedented scale.

- Internet Robustness[2], a collaboration between Harvard University, the Los Alamos National Laboratory, and Old Dominion University aimed at increasing link robustness by specifying how to express information about archival versions of resources that are linked from an HTML page. The project has close ties with the study pertaining to reference rot in legal citations [5] and the perma.cc[3] effort aimed at pro-actively archiving resources linked from legal literature.

1.5 Memento Tools

A wide range of Memento compliant tools is meanwhile available, and the tutorial will provide an overview of the most prominent server-side and client-side ones, including Global Open Wayback, SiteStory, Memento MediaWiki extension, Memento Time Travel for Chrome, and mcurl.

1.6 Memento at Work

The power of Memento’s time travel will be illustrated by means of demonstration of both production and experimental versions of Memento-related tools.

2. ACKNOWLEDGMENTS

During its initial years, the Memento project received funding from the Library of Congress. Hiberlink is funded by the Andrew W. Mellon Foundation.

3. REFERENCES

[1] Van de Sompel, H., Nelson, M. L., and Sanderson, R. 2013. RFC 7089: HTTP Framework for Time-Based Access to Resource States – Memento. http://ietf.org/rfc/rfc7089.txt

[2] Van de Sompel, H. 2013. Memento Guide: Resource Versioning and Memento. http://mementoweb.org/guide/howto/

[3] Van de Sompel, H., Klein, M., Sanderson, R., and Nelson, M.L. 2013. Thoughts on Referencing, Linking, Reference Rot. http://mementoweb.org/missing-link/

[4] Sanderson, R., Phillips, M., and Van de Sompel, H. 2011. Analyzing the persistence of referenced web resources with Memento. http://arxiv.org/abs/1105.3459

[5] Zittrain, J., Albert, K., and Lessig, L. 2014. Perma: Scoping and addressing the problem of link and reference rot in legal citations. http://www.harvardlawreview.org/issues/127/february14/forum_1031.php

[2] Internet Robustness, http://cyber.law.harvard.edu/research/internetrobustness

[3] perma.cc, http://perma.cc

iPRES MELBOURNE 8-10 OCTOBER

Digital Preservation Systems Showcase

Time: Tuesday, 7th October 2014
Venue: Theatrette

Description: This session explores the functionality of digital preservation systems available to the user community. System developers showcase their systems in line with a pre-determined set of functions. These functions are derived from current standards, key literature and interest groups (e.g. PREMIS, OAIS, and the International Internet Preservation Consortium Preservation Working Group).

Running order:

09:00 – 09:15	Introduction
09:15 – 10:15	DuraSpace
10:15 – 11:15	Artefactual Systems
11:15 – 11:40	Break
11:40 – 12:40	KEEP Solutions
12:40 – 13:40	Lunch
13:40 – 14:40	Preservica
14:40 – 15:00	Break
15:00 – 16:00	Ex Libris
16:00 – 17:15	Questions and round-up

Presentation structure

The DP system providers explored the questions listed below, demonstrating how their products handle these core components of digital preservation. The fundamental issue was to highlight how their systems tackle the key areas, taking into account the contextual 'Considerations' listed below.

Showcase details

The preservation workflow for digital preservation systems can be simplified into three large groups.

- How do we get content in?
- How do we manage and preserve it once in?
- How can the content be accessed from the system?

More detailed functional areas are listed under each of these groups.

How do we get it in?

- Ingest flows / methods

  What are the flows that can be used to route digital content into the system. Are there difference between flows (for example, different assessment criteria, more detailed identification, stricter security?)

- Preconditioning / pre-ingest preparation

Does the system take care of any actions that may be considered 'preparation' of content for ingest. This may include such actions as adding correct file extensions, repairing 'broken' files, tidying of file names.

- Format identification

  How does your system identify formats? To what level is identification made? What tools and resources does it use? How is format identification used by the system?
- Metadata extraction

  How does your system extract metadata? This may include technical and descriptive metadata. How much metadata is extracted and for what purpose? What tools and resources does it use?
- Fixity checking/assignation

  Does the system check fixity supplied to it? What type of fixity recording/checking mechanisms are used?
- Virus checking

  Does the system check for malware? What tools does it use? What happens if a virus is discovered?

How do we manage and preserve it?

- Intellectual management

  Is intellectual management (e.g. cataloguing) done by the system, or this a dependency on another system?
- Risk analysis

  Does the system do risk analysis of content based on technical form of the content? If so, what is it checking? What information is given to system users and what can they do with that information?
- Preservation planning

  What is the process for preservation planning in the system? What tests/proofs/sign-offs are required? How does it relate, if at all, to current community practice in (for example PLATO[1]*)?*
- Preservation execution

  How does your system undertake preservation actions (migration and/or emulation)?
- Repository management (queries, monitoring, analysis, updates)

  What reports is the system capable of generating? Is there a repository dashboard to analyse contents? What statistics are delivered out-of-the-box? Can functions such as format identification, virus checking, fixity checking and metadata extraction be re-run as required? How are updates for third party tools dealt with?
- Exception handling

  What functions exist for users to deal with exceptions in any process (including ingest processes?)

How do we access it?

- Derivative generation (static, on-the-fly, options of types)

  How is access given to content? Are access copies made of masters? If so, what are the formats used (and what are their master formats)? How are these copies made?
- Access rights

  How are rights administered and managed by the system?
- Complex materials

  Are there special access methods available for particularly complex materials such as email, webharvests, full text, multiple-object materials?
- Handing over to other access methods

  Can the system hand over materials to other access mechanisms? How easy is this handover? Are there any constraints?

[1] http://www.ifs.tuwien.ac.at/dp/plato/intro/.

- Export of data
  Can data be exported from the system?

Considerations

- Flexibility/interoperability of the system
  What external (to the system) sources is the system dependent on? How is reflected in the concept of the Archival Information Package? How are updates in those dependencies managed (for example content management system changes to access rights, or identifiers).
- Exit strategy
  How locked-in are customers?
- Archival Information Package
  - Relationship to PREMIS and other metadata schemas.
    Does the system implement PREMIS? Is it a conformant implementation? If not, why not?
  - Data model
    What is the object model used by the system? What level of detail captured about the object (intellectual entity only, or all the way down to bitstream information [as per PREMIS])?
- Provenance
  What data is kept to track/note provenance of the content? What triggers the new generation of metadata in this trail?
- Large/small, bulk/single
  How does the system deal with the very large and the very small and boutique (both in terms of size and number?
- Testing
  What testing regimes/tools are in place for new releases?
- Storage
  Does the system promote a particular type of storage? Are there any constraints on the configuration of storage?

Moderator

Moderated by Ross King, Chairman of the Board at Open Planets Foundation and Senior Scientist at the Austrian Institute of Technology.

iPRES MELBOURNE 8-10 OCTOBER

Digital Preservation Systems Showcase – Audience Notes

DuraSpace / Artefactual Systems / KEEP Solutions / Preservica / Ex Libris

How do we get it in?

Ingest flows / methods

……

……

……

……

Preconditioning / pre-ingest preparation

……

……

……

……

Format identification

……

……

……

……

Metadata extraction

……

……

……

……

Fixity checking/assignation

……

……

……

……

Virus checking

……

……

……

……

How do we manage and preserve it?

Intellectual management

……

……

……

……

Risk analysis

……

……

……

……

Preservation planning

……

……

……

……

Preservation execution

……

……

……

……

Repository management (queries, monitoring, analysis, updates)

……

……

……

……

Exception handling

……

……

……

……

How do we access it?

Derivative generation (static, on-the-fly, options of types)

………

………

………

………

Access rights

………

………

………

………

Complex materials

………

………

………

………

Handing over to other access methods

………

………

………

………

Export of data

………

………

………

………

Considerations

Flexibility/interoperability of the system

………

………

………

………

Exit strategy

………

………

………

………

Archival Information Package

……
……
……

- ○ Relationship to PREMIS and other metadata schemas.

……
……
……

- ○ Data model

……
……
……

Provenance

……
……
……
……

Large/small, bulk/single

……
……
……
……

Testing

……
……
……
……

Storage

……
……
……
……

Modelling File Formats and Technical Environments using the NSLA Digital Preservation Technical Registry (DPTR)

Jay Gattuso, Peter McKinney, Steve Knight
National Library of New Zealand
Cnr Molesworth & Aitken St
Wellington
New Zealand
Jay.Gattuso@dia.govt.nz

Jan Hutař, Ross Spencer
Archives New Zealand
10 Mulgrave St
Wellington
New Zealand
Jan.Hutar@dia.govt.nz

Libor Coufal
National Library of Australia
Parkes Place
Canberra ACT 2600
Australia
lcoufal@nla.gov.au

Kevin DeVorsey
National Archives and Records Administration
One Bowling Green, 3rd Floor
New York, NY 10004
USA
Kevin.DeVorsey@nara.gov

ABSTRACT

This workshop introduces the work of the National and State Libraries funded work on a Digital Preservation Technical Registry. In particular it will allow participants to gain an understanding of the new model for modelling formats. They will be tasked with working through exercises designed not only to give participants an understanding of the model, but to test and critique it.

General Terms

infrastructure, communities, strategic environment, specialist content types,

Keywords

Technical Registry, Models, File Formats, Hardware, Software, Community, Collaboration.

1. PROPOSAL

The heart of contemporary digital preservation is multi-faceted. Where once the technical registry sat at its core, we now seek 'war stories' from the community that detail experiences with legacy digital information. We seek high quality information about file formats, carrier mediums, software, and the complete picture of the technical environments that we're dealing with. With archival principle at the centre of our work we also look for detailed provenance, validation and verifiability of that information.

Along with that, the technical registry still holds a key role in the community's multi-faceted approach. The technical registry, through enabling file format identification and validation, can aid the filtering and routing of content at the pre-deposit, and pre-ingest stages of the digital preservation lifecycle.

The technical registry is still the core information source for the migration and maintenance of content as part of any technology-watch capability or preservation action.

We believe that the technical registry should provide information that is accessible to all involved in digital preservation at all levels of skills and knowledge. The information also needs to be actionable, that is, machine readable information that can be accessed by the tools in the digital preservation toolkit. Registries should use relationships to describe more complete technical environments – the links between specific instances of software, hardware, carrier mediums and file formats. Registry information also needs to be augmented with user-level, 'community text' describing war-stories, domain expert knowledge, and the institutional relevance of specific registry entries. Our aim is to create a registry flexible enough to contain and identify real life file format instances with all their specialities and varieties.

Striking that balance, we'll be introducing our work on the NSLA Digital Preservation Technical Registry, providing an overview of its core features before introducing our more radical changes in thinking. The biggest advancement we'll introduce is the overhaul of the format model traditionally used in digital preservation - presenting three interpretations of file format that we believe encompass the many different ways we talk about the subject in the community. Workshop participants will be given a more thorough introduction to this side of the work package, learning about these three components and the building blocks used to create them – format Aspects.

Using Aspects and the knowledge of how to create our three format objects, participants will engage in a modelling activity to help us challenge the work we've completed this far and help us to reinforce a step-change in thinking about the requirements of a modern, comprehensive technical registry.

2. DRAFT AGENDA

0900-1000: Introduction to the NSLA Digital Preservation Technical Registry
1000-1100: The new format paradigm
1100-1115: Coffee and refreshments
1115-1230: Breaking down a format specification
1230-1330: Lunch
1330-1500: Building an Implementation

1500-1515: Working across the format domain
1515-1645: Use case and the wider Registry environment
1645-1700: Wrap up

3. WHO SHOULD ATTEND?

Digital archivists, digital preservation analysts and developers.
Digital preservation service providers and organisations.
Repository managers and organisational digital preservation, information technology leadership.

4. WORKSHOP OUTCOMES

The primary purpose of the workshop was to introduce the format model that the NSLA Digital Preservation Technical Registry team has been developing. The workshop was attended by a wide-spread of people who were eager to hear about the Technical Registry work. The beginning of the session was side-tracked slightly away from the format model specifically, but this, it runs out was necessary to explain the context of the format work. One learning from this was the need for the team to spend more time setting up the reasons for the Registry and the use cases that it fulfills.

The outcomes from the workshop are shaped by the scope of the event. The workshop was a ‘transmit’ event. That is, it was the first time that the (majority) of participants had been introduced to the format work that had been developing over the past three years. The goal was therefore to successfully describe the format model in order that the participants could not only give initial thoughts but more importantly, spend time considering the work and engage in deeper discussions at a later date.

Essentially, the team believe that the participants felt that the work was going down the right track. It will be hard to achieve and there are still a number of areas unresolved and questions unanswered, but these were points that the team were aware of (and in many cases, there was not time to satisfy all questions, particularly those that were about the broader project rather than the format model).

.

Acquiring and Processing Born-Digital Data Using the BitCurator Environment

Christopher A. Lee
School of Information and Library Science
University of North Carolina
216 Lenoir Drive, CB #3360
1-(919)-966-3598
callee@ils.unc.edu

ABSTRACT

This tutorial will prepare participants to use the open-source BitCurator environment to acquire and process born-digital data. There will be a brief lecture and discussion that focuses on the motivation for using the tools and several foundational technical concepts. The remainder of the tutorial will be devoted to demonstration and hands-on exercises that demonstrate specific tools and methods. Participants will learn how to mount media as read-only, create disk images, mount forensically packaged disk images, export individual files or entire directories from disk images, use Nautilus scripts to perform batch activities, generate and interpret Digital Forensics XML (DFXML), generate a variety of standard and customized reports (including PREMIS recods), and identify various forms of sensitive data within collections.

Categories and Subject Descriptors

H.3.7 [**Information Storage and Retrieval**]: Digital Libraries – *collection, dissemination, systems issues.*

General Terms

Provenance, Data Triage, Digital Forensics.

Keywords

Forensics, preservation, DFXML, metadata, privacy, collections, acquisition

1. BITCURATOR PROJECT

The BitCurator Project, a collaborative effort led by the School of Information and Library Science at the University of North Carolina at Chapel Hill and Maryland Institute for Technology in the Humanities at the University of Maryland, is addressing two fundamental needs and opportunities for collecting institutions: (1) integrating digital forensics tools and methods into the workflows and collection management environments of libraries, archives and museums and (2) supporting properly mediated public access to forensically acquired data [4].

2. BITCURATOR ENVIRONMENT

We are developing and disseminating a suite of open source tools. These tools are being developed and tested in a Linux environment; the software on which they depend can readily be compiled for Windows environments (and in most cases are currently distributed as both source code and Windows binaries). We intend the majority of the development for BitCurator to support cross-platform use of the software. We are freely disseminating the software under an open source (GPL, Version 3) license. BitCurator provides users with two primary paths to integrate digital forensics tools and techniques into archival and library workflows.

First, the BitCurator software can be run as a ready-to-run Linux environment that can be used either as a virtual machine (VM) or installed as a host operating system. This environment is customized to provide users with graphic user interface (GUI)-based scripts that provide simplified access to common functions associated with handling media, including facilities to prevent inadvertent write-enabled mounting (software write-blocking).

Second, the BitCurator software can be run as a set of individual software tools, packages, support scripts, and documentation to reproduce full or partial functionality of the ready-to-run BitCurator environment. These include a software metapackage (.deb) file that replicates the software dependency tree on which software sources built for BitCurator rely; a set of software sources and supporting environmental scripts developed by the BitCurator team and made publicly available at via our GitHub repository (links at http://wiki.bitcurator.net); and all other third-party open source digital forensics software included in the BitCurator environment.

3. TUTORIAL FORMAT

This is being proposed as a full-day (6-hour) format. There will be a brief lecture and discussion that focuses on the motivation for using the tools and several foundational technical concepts. The remainder of the tutorial will be devoted to demonstration and hands-on exercises that demonstrate specific tools and methods.

4. INTENDED AUDIENCE

This tutorial should be of interest to information professionals who are responsible for acquiring or transferring collections of digital materials, particularly those that are received on removable media. Another intended audience is individuals involved in digital preservation research, development and IT management, who will learn how data generated within the BitCurator environment can complement and potentially be integrated with data generated by other tools and systems.

5. EXPECTED LEARNING OUTCOMES

This tutorial will prepare participants to use the open-source BitCurator environment to acquire and process born-digital data. Tools that BitCurator is incorporating include Guymager, a program for capturing disk images; bulk extractor, for extracting features of interest from disk images (including private and individually identifying information); fiwalk, for generating Digital Forensics XML (DFXML) output describing filesystem hierarchies contained on disk images; The Sleuth Kit (TSK), for viewing, identifying and extraction information from disk images; Nautilus scripts to automate the actions of command-line forensics utilities through the Ubuntu desktop browser; and sdhash, a fuzzing hashing application that can find partial matches between similar files. For further information about several of these tools, see [1,2,3,5].

Upon completion of this tutorial, participants should understand several of the major motivations and uses cases for applying the BitCurator environment. They will also know how to perform the following tasks:

- mount media as read-only
- create disk images, mount forensically packaged disk images
- export individual files or entire directories from disk images
- use Nautilus scripts to perform batch activities
- generate and interpret Digital Forensics XML (DFXML) generate a variety of standard and customized reports (including PREMIS records)
- identify various forms of sensitive data within collections.

Participants will also become aware of the resources that are available for learning more about the software and engage with other users after completion of the tutorial.

6. INSTRUCTOR BIOGRAPHY

Christopher (Cal) Lee is Associate Professor at the School of Information and Library Science at the University of North Carolina, Chapel Hill. He teaches graduate and continuing education courses in archival administration, records management, digital curation, and information technology for managing digital collections. His research focuses on curation of digital collections and stewardship of personal digital archives. Cal is PI for the BitCurator project and editor of *I, Digital: Personal Collections in the Digital Era*.

7. ACKNOWLEDGMENTS

BitCurator development has been supported by the Andrew W. Mellon Foundation. Members of the BitCurator team are Alexandra Chassanoff, Matthew Kirschenbaum, Christopher (Cal) Lee, Sunitha Misra, Porter Olsen, and Kam Woods. Members of two advisory boards have made valuable contributions: the Development Advisory Group (DAG) and Professional Experts Panel (PEP).

8. REFERENCES

[1] Cohen, M., Garfinkel, S., and Schatz, B. 2009. Extending the Advanced Forensic Format to Accommodate Multiple Data Sources, Logical Evidence, Arbitrary Information and Forensic Workflow. *Digital Investigation* 6 (2009), S57-S68.

[2] Garfinkel, S. Digital Forensics XML and the DFXML Toolset. *Digital Investigation* 8 (2012), 161-174.

[3] Garfinkel, S.L. Providing Cryptographic Security and Evidentiary Chain-of-Custody with the Advanced Forensic Format, Library, and Tools. *International Journal of Digital Crime and Forensics* 1, 1 (2009), 1-28;

[4] Lee, C.A., Kirschenbaum, M.G., Chassanoff, A., Olsen, P., and Woods, K. BitCurator: Tools and Techniques for Digital Forensics in Collecting Institutions. *D-Lib Magazine* 18, 5/6 (May/June 2012).

[5] Roussev, V. An Evaluation of Forensic Similarity Hashes. *Digital Investigation* 8 (2011), S34-S41.

Applying the TIMBUS Approach to Preserving Context in Digital Libraries

Carlos Coutinho
Caixa Magica Software Rua Soeiro Pereira Gomes, Lote 1-4°B Edifício Espanha, 1600-196 Lisboa +351 217 921 260
carlos.coutinho@caixamagica.pt

Paul Gooding
Digital Preservation Coalition
British Library, Floor 5, Room 14
96 Euston Road, London, NW1 2BD +44 (0) 20 7412 7329
paul@dpconline.org

ABSTRACT

To date, digital preservation has generally focused on the preservation of specific data in the form of artefacts. However, in order to ensure that this data is accessible in the long term, it is vital that we consider how to extract and preserve information on the software and hardware contexts which this data depends upon to operate. We therefore need tools to assist in identifying, preserving and validating the processes which underpin the creation of data in digital libraries.

In particular, we need to consider the importance of preserving not just individual digital artefacts, but the platforms which allow digital libraries to render or execute their items. Digital libraries rely on this software to render their items, and it is therefore important to know configuration details and software dependencies to ensure these items remain fully operational in the future. In the case of digital libraries, the TIMBUS framework provides the tools necessary to assist practitioners in identifying relevant processes, undertake risk analysis, and then to assist the user in extracting, preserving and revalidating the necessary processes.

This half-day tutorial introduces the TIMBUS approach to process preservation, and demonstrates how it can be applied to issues relating to digital libraries. TIMBUS focuses primarily on business processes, but this tutorial will show its approach to process-oriented preservation is also relevant to digital libraries. It provides a methodology for process preservation and a set of tools which help to semi-automatically validate and preserve processes so that they can be recreated at a later dat. Participants will be given the knowledge to understand the importance of technical environments for collection items, and learn more about the TIMBUS solutions through examples relevant to the digital library domain. They will also gain an understanding of digital preservation as a risk mitigation strategy.

1. TUTORIAL LEVEL

Introductory level.

2. DURATION

Half-day (3 hours).

3. PARTICIPANT NUMBERS

Up to 20.General Terms

4. OUTLINE OF THE CONTENT

The tutorial will cover the following topics:

- An introduction to the basics of digital preservation: here the purpose is to present an overview of the state-of-the-art in digital preservation techniques and motivations, and to equip participants with the requisite knowledge for the tutorial;
- Discussion of moving from data-oriented preservation to process-oriented preservation: this section aims to show the problems regarding current approaches to business preservation. Whereas up to now the main concern was to capture the business data (e.g., databases and logs), the paradigm is shifting to understanding the underlying business processes and, besides modelling and capturing, to preserve the complete surrounding environment;
- Explanation of the TIMBUS approach to process and context preservation by presenting the TIMBUS storyboard;
- Introduction to the TIMBUS set of tools and showcase of the TIMBUS architecture model, including:
    - Context Capturing tools;
    - Context Model tools;
    - Risk Management tools;
    - Digital Preservation Expert Suite, Preservation Identifier and Dependencies Reasoner;
    - The TIMBUS methodology for process preservation;
- Presentation of a Context Model for capturing and describing processes: this will define the context model ontologies (Domain Independent Ontologies and Domain-Specific Ontologies) and corresponding support tools, including Archi, Archi2OWL, Jena, Protégé;
- Discussion of the challenges of automatic and semi-automatic capture of context, including definition of the context capturing tools;
- Explanation of how to adapt tools and models to the heterogeneity of systems and businesses, including a showcase of the paradigm of development by

contributions and architecture off the context metadata capturing tools using OSGi;

- Explanation of process capture and modelling from real business artefacts, including definition of tools used to capture and model business processes based on real business evidence sources such as logs;
- Introduction to variants of risk analysis: Classical and Simulation. The suite of risk management tools in TIMBUS will be introduced, including their features and approaches towards:
    - Classical Risk Management: Risk Evaluation and Treatment tool;
    - Simulation Risk Management: Intelligent Enterprise Risk Management tool (iERM);
- Explanation of Digital Preservation as a risk management strategy. This will present the Digital Preservation Expert Suite of tools for performing Digital Preservation as one solution provider for Risk Management mitigation.

5. INTENDED AUDIENCE

The tutorial is aimed at researchers, publishers and curators in digital libraries, who want to learn about process-oriented digital preservation. Some understanding of digital preservation will be helpful, but the tutorial will begin with an introduction to the basics.

6. EXPECTED LEARNING OUTCOMES

The tutorial participants will gain understanding of the concept and importance of process-oriented preservation. They will learn how to adapt tools and models for this purpose, how to capture processes, and the basics of risk analysis for process preservation. They will learn more about the TIMBUS suite of tools, and be introduced to the role of these tools in successfully preserving the entire business environment, including both data and business-oriented preservation. As a result, participants will be empowered to explore these issues in relation to their own organisations.

7. BIOGRAPHY OF THE PRESENTER(S)

Carlos Coutinho is a Senior Research engineer and R&D Project Manager at Caixa Mágica Software in Lisbon Portugal. He holds a PhD degree in Electrical and Computer Engineering (2013), awarded by the New University of Lisbon (FCT-UNL), Portugal, where he also does research, with interests in Enterprise Interoperability, Adaptable Platforms and Systems, SOA, and Model-Driven Engineering. He has more than ten years of experience teaching the fields of IT at Portuguese universities ISEL, ISCAL, ISGB and ISCTE. He has five publications in international scientific journals (with ISI-IF) and more than fifteen publications in peer-reviewed international conferences, and is part of the review committee of two journals and three yearly conferences. He also holds a PMI-PMP© title and has a post-graduation in Project Management by Instituto Superior Bissaya-Barreto (ISBB) in Coimbra, Portugal. He has more than fifteen years of experience working as an engineer in the enterprise IT area, working in several fields from ICT, Services, Public Administration and the Aerospace industry, in several multinational companies like Alcatel, Siemens and Critical Software.

Other presenters to be confirmed.

8. Workshop Outcomes

- A very important outcome that was intended with this workshop was for the TIMBUS development team to have a clear feel of the impact of the project's tools on the final customers and on the potential users that are going to support the project outcomes. More than producing outputs that fill the objectives of the project on the EC reviews, TIMBUS intends to produce tools that are in fact usable for DP customers and practitioners;
- The community that assisted the workshop was indeed quite interested in the outcoming tools that were produced and showcased, particularly on the context model ontologies and on the business process extraction framework, as well as on the whole concept of risk-driven digital preservation.

Surveying ISO Standards for PDF: archive, accessibility, engineering, metadata, 3D data and PDF itself. History, pain points, solutions and the dream.

Duff Johnson
PDF Association
Neue Kantstrasse 14
14057 Berlin, Germany
+1 617 283 4226
duff.johnson@pdfa.org

ABSTRACT

This workshop provides high-level information on the features, development and current status of ISO standards for PDF relevant to archiving. It then provides an opportunity for participants to clearly identify and prioritize related requirements and solutions.

Presented by the ISO Project Leader of ISO 32000, an independent consultant, the workshop will begin by reviewing the features and history of the technology and its associated standards. Thereafter, a set of breakout sessions will challenge participants to clearly specify their ideas and concerns about PDF and suggest a variety of solutions leveraging commercial software interests.

Finally, workshop participants will come together to discuss the breakout session results, identify common themes and build consensus on specific, actionable requests for PDF producer and processor developers in the short and medium term.

General Terms

strategic environment, preservation strategies and workflows, specialist content types, digital preservation marketplace

Keywords

PDF, PDF/A, PDF/E, PDF/UA, XMP, PRC, software, standards

1. OUTLINE

Following a series of short presentations introducing ISO-standardized PDF technology, the balance of time is occupied by breakout sessions followed by a group discussion.

1.1 Presentations (80 minutes)

This segment begins with a review of the procedures to be followed in the workshop together with a brief review of the proposed breakout session questions.

Next the workshop will hear a set of short presentations surveying PDF features, history and standards development, with ample time for questions. The segment closes with a group discussion in which three topics are selected for the breakout sessions.

The presentation segment covers:

1) workshop organization, release form;
2) more than rendering: PDF's value proposition & feature-set
3) ISO Standards for PDF: File format standards of interest to digital preservation professionals, including:
    a. ISO 32000 parts 1 and 2 (PDF 1.7 and PDF 2.0);
    b. PDF/A parts 1, 2 and 3;
    c. PDF/E parts 1 and 2;
    d. PDF/UA;
    e. XMP (PDF's XML-based metadata standard);
    f. PRC (PDF's ISO-standardized 3D file-format).
4) PDF history: from Adobe to ISO to PDF 2.0;
5) the PDF future: standards development and the marketplace;
6) review some options, then create the breakout session topics.

1.2 Break (10 minutes)

A 10 minute break.

1.3 Breakout sessions (90 minutes)

This segment provides a means for workshop participants to collaborate on sketching a common understanding of key pain points related to some controversial, technical, conceptual or other participant-chosen topic specific to PDF technology.

Just before the break the workshop participants selected three breakout session topics. Following the break participants choose and join one of the three breakout sessions. After 30 minutes the groups come back together and present each breakout session's results to the rest of the workshop participants for review.

In addition to reviewing and characterizing these concerns, participants will be asked to propose specific actions they'd like to see industry take to address them.

The objective of each breakout session may be stated thusly:

1) 5 minutes for online research or just focused thinking;
2) 1 minute to select a minute-taker and facilitator
3) 24 minutes to discuss and achieve consensus on:
    a. **Up to 3 key pain points or big opportunities** identified, characterized, ranked and rated for significance to digital preservationists;
    b. **3 conceivable market-driven solutions** (conservative, plausible, shoot-the-moon) that would address one (or more) pain points;
    c. **3 key reasons to be pessimistic or optimistic** about each proposed solution. If there's time, address both perspectives.

Following discussion, and after the workshop ends, the breakout session results will be anonymized, and distributed to participants and disseminated within the industry.

1.3.1 Breakout session topics

With a target of 3-4 breakout sessions, each session's topic is decided by a vote of all workshop participants. During the break, participants choose which session to attend.

Centered on PDF technology, sessions may be entirely technical, practical, theoretical, marketplace-specific, or otherwise. The presenter will visit each breakout session to provide feedback, quick answers (if available) or other assistance.

1.3.2 Breakout topic examples

- PDF/A-3: The context, challenge and opportunity of embedded non-archival formats
- The roles and responsibilities of PDF/A processors
- Should there be a PDF/A-4? What should it look like?
- Are we missing any conformance levels from PDF/A?
- Archiving dynamic data: is PDF/E the right way?
- Assume PDF 2.0 is great, with no ambiguities... what should PDF 3.0 look like? Or should we not bother?
- Why TIFF when there's PDF?
- Why PDF when there's EPUB? Or vice-versa
- What's PDF worth when Acrobat includes text-editing
- The PDF icon Adobe products install is actually the Adobe Acrobat icon. No wonder users still call PDF files "Adobe"! Does it matter? Discuss
- How could conformance level "a" become useful?
- "Enough with this theory and standards: 80% of my problems with PDF are due to X. Let's talk about X!"
- What should a PDF/A validator do, exactly? Just syntax, or rendering too? Certification? Does PDF validation (as opposed to PDF/A) matter to you, and why?
- Is PDF a nearly-perfect set of features for digital preservation of textual content? What's it missing?

2. INTENDED AUDIENCE

Archives policy-makers and practitioners, records-managers, ECM implementers.

3. EXPECTED LEARNING OUTCOMES

Tutorial attendees will gain authoritative high-level knowledge of the archival-relevant PDF standards and respective conformance levels. They will be invited to experience and participate in a product-management exercise conducted from the industry perspective. Participants will gain deeper understanding of the commercial view of PDF technology from exposure to industry and standards-development orientations. They will learn how to access industry-sponsored activities and resources and participate in PDF standards development and best-practices efforts.

4. PRESENTER'S BIOGRAPHY

Duff Johnson began his career with PDF in 1996 when he founded an electronic document service bureau dedicated to PDF technology-based solutions. Having managed three companies in the PDF technology space, he is now an independent consultant.

Johnson began working in PDF standards development in 2005 as chairman of the AIIM committee developing PDF/UA, now ISO 14289. Elected Project Leader for ISO 32000 in 2011, Johnson chairs the committee developing PDF 2.0, the next generation of the world's chosen fixed-layout electronic document format.

Johnson serves as Vice Chairman of the PDF Association (www.pdfa.org), the vendor organization for the worldwide PDF software industry. He is currently serving as Standards Committee Chairman and as a member of AIIM's Board of Directors.

Leveraging Web Archiving Tools for Research and Long-Term Access

Lori Donovan
Internet Archive
300 Funston Ave
San Francisco, CA 94118
1-415-561-6799 x4
lori@archive.org

ABSTRACT

This workshop will introduce participants to web archiving concepts and challenges, including creating web archives and providing for access and research use.

General Terms

Preservation strategies and workflows, specialist content types, training and education.

Keywords

Web archiving, research services, access

1. INTRODUCTION

Web archiving is an important part of the digital preservation field. While most are familiar with the Wayback Machine available at archive.org, less are aware that there are a number of tools and services developed for organizations and individuals to create their own web archives, including the capability to search and analyze large data sets built around the WARC file format, an ISO standard for web archiving. In addition, web archives provide permanent URLs for citation and can show how a website has changed over time at a single URL, even if no longer available on the live web. In short, web archives can provide very necessary preservation tools for researchers and archivists to manage content that is only posted on the web.

This workshop will introduce participants (15--20) to basic web archiving concepts and challenges. Using the Archive--It (www.archive--it.org) web application, participants will have a hands--on opportunity to build a collection of content archived from the web, which can include their own organization's web presence, social media, digital exhibitions, data sets, or topical content publicly available on the web. Following the workshop participants will have a searchable archive available to them, including the option of downloading WARC files for long-term preservation or research.

The target audience for this workshop includes interested scholars researching the web and professionals responsible for digital library services or digital archives. No prerequisite knowledge of or experience with web archives is necessary, and the session does not require any programming or advanced technical knowledge of the web. The workshop will not be oriented towards those with deep knowledge of web archives or the WARC format, although there could be time allotted to a demonstration of another web archiving tool or project related to web archiving and this should be specified in the CFP (see below).

2. PARTICIPANT INFORMATION

In order to make the most of the workshop and ensure that the curriculum is tailored toward participant interest, some additional information about participants would be helpful. It should include:
--Description of participant interest areas and/or professional projects.
--Description of prior experience with using web archives or their own web archiving (if applicable).
--5 to 10 websites to be archived as part of 1 or more collections of content, and links to the Robots.txt files if applicable. More information is here:
https://webarchive.jira.com/wiki/display/ARIH/Robots+Exclusion+Protocol
With permission from participants, URLs will be crawled as a test (no data archived) prior to the workshop so post crawl reports can be analyzed as part of the workshop curriculum.
If possible, the instructor should receive this information at least 1-2 weeks before the workshop.

3. ABOUT THE INSTRUCTOR

Lori Donovan is a Partner Specialist at the Internet Archive helping libraries, museums and other cultural institutions archive web content. Over the past four years, Lori has given more than 25 presentations on web archiving at library, archives and digital preservation conferences both in the United States and internationally. Lori has a Masters of Science in Information from the University of Michigan specializing in Archives and Digital Preservation.

Posters and Demonstrations

Functional Access to Electronic Media Collections using Emulation-as-a-Service

Thomas Bähr, Michelle Lindlar
Technische Informationsbibliothek
Hannover, Germany
firstname.lastname@tib.uni-hannover.de

Klaus Rechert and Thomas Liebetraut
University of Freiburg
Freiburg, Germany
firstname.lastname@rz.uni-freiburg.de

ABSTRACT

Over the last 30 years the German National Library of Science and Technology (TIB) accumulated a large collection of various electronic media, such as floppies or CD-ROMs. This poster describes both practical workflows as well as technical infrastructure to provide authentic and interactive access to the TIB's large electronic media collection.

General Terms

Case Studies and Best Practice

Keywords

Emulation, Access, Media Collection

1. CURATION CHALLENGES

The TIB – German National Library of Science and Technology – is the national subject library for all areas of engineering, architecture, chemistry, information technology, mathematics and physics. The library provides national and international research and industry with information regardless of the information's language or material type. It furthermore functions as a "library of last resort" for the specified subject areas and has a legal mandate for archiving. As an archival library, the TIB has dedicated staff and resources for digital preservation activities.

As part of the digital preservation activities, the TIB is currently analyzing it's holdings on removable data carriers. Here, the cataloguing practice of the past 30 years proves to be problematic when investigating exact numbers of items by carrier type per collection, due to the fact that any electronic source was often described simply as "electronic media" in the catalogue, thus lacking a distinction between, e.g., CD-ROM, CD-R, DVD, floppy or online source. A first analysis of a few selected collections has brought forth estimates for optical data carriers (CD-ROM, CD-R, CD-RW, DVD) as summarized in Table 1. A sampling of the evaluated collections established that the content of the data carriers is often complex: often, a carrier may contain a combination of software needed to render the file as well as the digital object itself.

Table 1: TIB optical media inventory

Collections	Media Items
Supplements to Monographs	∼ 20.000
Patents, Rules and Standards	∼ 18.000
Conference Proceedings	∼ 14.000
Serials	∼ 13.000
Lexica, Dictionaries, Databases	∼ 300

The optical carrier is still a requested medium – the supplements to monographs, for example, which can be only viewed within the library's reading rooms, booked around 1.500 requests in 2013. A different statistic showed that in the month of January of 2014 alone, a total of 2.819 pages of inter-lending requests fulfilled by TIB were generated from information that the library held only on CD carriers.

In the light of deteriorating data carriers as well as hardware and software dependencies of the materials contained on the data carriers, preservation and continuous access strategies for this material type need to be developed and implemented. While bit preservation issues can be addressed by moving the information off of the original carrier – either in form of ISO9660 images and/or in form of bitwise copies – the logical preservation activities, especially for complex objects held on the data carriers, is a different matter. One potential way to preserve the accessibility of the content with a high level of authenticity and utility is by using emulation. To test the feasibility of this approach, the TIB is currently implementing a pilot workflow with the University of Freiburg based on the "Emulation-as-a-Service" infrastructure.

Emulation-as-a-Service

Until now emulation has been seen as domain reserved for technical experts. Furthermore, emulation did not scale well due to the laborious preparation and technical setup procedures. Driven by the principles of division of labor and based on the observations on potential stakeholders a scalable service model has been developed – Emulation as a Service (EaaS) [1]. EaaS provides a modular set of technical building blocks (*emulation components*) to standardize deployment and to hide individual emulator complexity. Each emulation component encapsulates a specific emulator type, i.e. an emulator capable of replicating a certain system architecture, as an abstract component with a unified set of software interfaces (API). This way, different emulators can be integrated and can be used in dedicated archival workflows. For this purpose, EaaS offers users various options to interact interactively with an emulated environment, e.g. through remote desktop protocols (VNC, RDP) or, more conveniently, through a HTML5-enabled web browser. Furthermore, emulation components can be dynamically deployed in a large-scale cluster or Cloud infrastructure upon request. Hence, no spare computing resources have to kept available. An EaaS service-provider then is responsible for efficient hardware utilization and concentration of technical expertise and thus lighten the memory institutions' technical workload and requirements on necessary infrastructure.

2. WORKFLOWS

On the library side of the workflow, a number of pre-ingest steps need to be conducted. The legacy CD collection needs to be evaluated to be able to prioritize data carrier migration. This may include a number of factors, such as evaluating the uniqueness of the collection, the age of the data carrier or its risk of being damaged and copyright clearance. In a next step, the content of the carrier needs to be replicated, e.g. by creating an ISO9660 image. The TIB will ingest the images into its digital preservation system, which functions as a bit preservation layer and keeps necessary metadata.

To secure long-term access to the media's content, its information object is to be prepared by using EaaS ingest workflows. Through these workflows it is possible to create or to modify emulation environments, i.e. an emulated hardware system, an operating system and software required to render the digital object. In a second step a digital object then is linked to a specific environment that is able to render this specific object. While preparing a rendering environment process is optional (a ready-made standard environment could be used), linking an environment and a digital object results in technical meta-data with an exact description of the environment's view-path and its configuration such that a deterministic re-enactment of the system and object becomes possible. During ingest, the EaaS workflow allows to test-run the environment. This allows for an evaluation of the rendering quality and performance of object and environment. Fig. 1 shows an example of this process.

In many cases, pre-configured standard environments are not sufficient to render a digital object. A preliminary evaluation of a sample set of digital objects provided by the TIB showed that almost all of the tested objects require proprietary multimedia frameworks that were usually not included with the operating system. Therefore, before these objects can be used, these frameworks or respective viewer applications have to be installed.

For instance, the object shown in Fig. 1 is an interactive training program for Microsoft Excel set in a "futuristic virtual teaching room." This training program requires the Video for Windows multimedia framework and uses a proprietary viewer application to render the interactive content. In order to make this object accessible, a base system has been selected (Windows 98 SE) and additional software installation steps were performed using the "setup.exe" installer program provided on the object's CD-ROM. After these installation steps, the modified environment has been archived, creating a derivative of the base system specifically designed for this specific object. The derivative consists only of the actual changes to the installation medium (the system hard disk) together with a stable reference (HDL) to the original base image, both to save storage space and to allow for distributed data management, i.e. store derivatives together with the object in a single repository. Output of the EaaS ingest workflow is an *emulation environment* description, defining the EC configuration and referencing both the prepared rendering environment and digital object. In our prototypical implementation we make use of the Handle system as persistent identifiers for both environments and objects.

To access the prepared object, the *emulation environment* description is used by an EaaS provider to allocate and setup a suitable EC. External data-sources, i.e. a specific software environment and the digital object itself, are resolved and attached to the EC. The user is then able to connect to and interact with the environment using an HTML5-enabled web browser. This allows for instant rendering of complex objects by clicking on the corresponding link when browsing the library's web-based catalogue. On the end library user layer, the link to the emulated object in its respective environment will be linked from the library catalogue. Access modalities will follow the traditional CD workflow and will only be available from within the library's reading room.

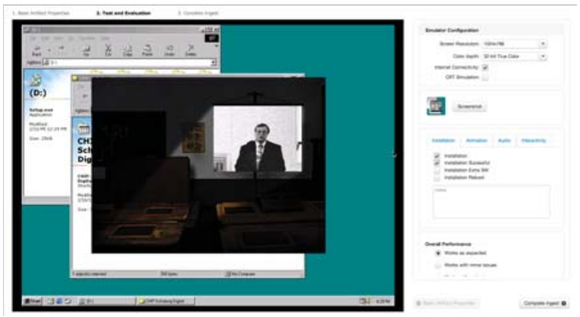


Figure 1: Assessment of a CD-Rom's content.

3. OUTLOOK

The current prototypical workflows address current curation challenges. Using EaaS provides novel access options to a significant part of the libraries electronic media collection. While the current state is not production ready yet, practical experience has been collected. The next steps, focus on further automation, e.g. automatically determine a suitable rendering environment for uniform object classes. For this, the sample will be significantly extended by TIB. On the library side, a workflow will be established to check the question of intellectual property rights for different subcollections in order to decide whether the EaaS for a specific group of works may only be offered within the reading room. Furthermore, the library plans to extend the EaaS workflow to the emulation of other data carriers, such as for example USB sticks, which are now starting to enter the collection especially in the area of congress and grey literature deposits.

4. REFERENCES

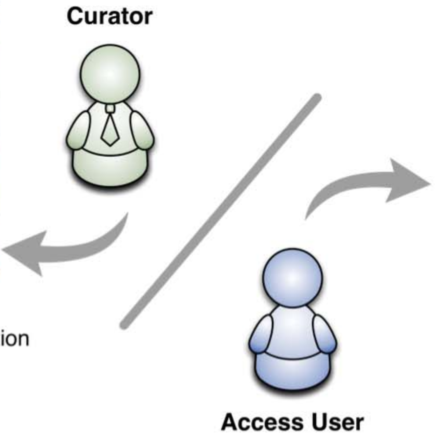
[1] T. Liebetraut, K. Rechert, I. Valizada, K. Meier, and D. von Suchodoloetz. Emulation-as-a-Service – The Past in the Cloud. In *7th IEEE International Conference on Cloud Computing (IEEE CLOUD)*, page to appear, 2014.

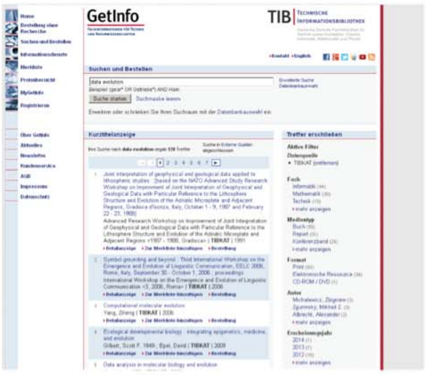
Functional Access to Electronic Media Collections using Emulation-as-a-Service

Thomas Bähr[1], Michelle Lindlar[1] und Klaus Rechert[2]

[1] German National Library of Science and Technology (TIB)
[2] University of Freiburg, Germany

User-Layer





GetInfo

TIB | TECHNISCHE INFORMATIONSBIBLIOTHEK

Home
Bestellung ohne Recherche
Suchen und Bestellen
Informationsdienste
Merkliste
Preisübersicht
MyGetInfo
Registrieren
Über GetInfo
Aktuelles
Newsletter
Kundenservice
AGB
Impressum
Datenschutz

Suchen und Bestellen

data evolution

Beispiel (gear* OR Getriebe*) AND Hain

Suche starten | Suchmaske leeren

Erweiterte Suche
Datenbankauswahl

Erweitern oder schränken Sie Ihren Suchraum mit der Datenbankauswahl ein.

Kurztitelanzeige

Ihre Suche nach data evolution ergab 539 Treffer

1 Joint interpretation of geophysical and geological data applied to lithospheric studies : [based on the NATO Advanced Study Research Workshop on Improvement of Joint Interpretation of Geophysical and Geological Data with Particular Reference to the Lithosphere Structure and Evolution of the Adriatic Microplate and Adjacent Regions, Gradisca d'Isonzo, Italy, October 1 - 9, 1987 and February 22 - 23, 1988]
Advanced Research Workshop on Improvement of Joint Interpretation of Geophysical and Geological Data with Particular Reference to the Lithosphere Structure and Evolution of the Adriatic Microplate and Adjacent Regions <1987 - 1988, Gradisca> | TIBKAT | 1991
› Detailanzeige › Zur Merkliste hinzufügen › Bestellung

2 Symbol grounding and beyond : Third International Workshop on the Emergence and Evolution of Linguistic Communication, EELC 2006, Rome, Italy, September 30 - October 1, 2006 ; proceedings
International Workshop on the Emergence and Evolution of Linguistic Communication <3, 2006, Roma> | TIBKAT | 2006
› Detailanzeige › Zur Merkliste hinzufügen › Bestellung

3 Computational molecular evolution
Yang, Zheng | TIBKAT | 2006
› Detailanzeige › Zur Merkliste hinzufügen › Bestellung

4 Ecological developmental biology : integrating epigenetics, medicine, and evolution
Gilbert, Scott F. 1949- ; Epel, David | TIBKAT | 2009
› Detailanzeige › Zur Merkliste hinzufügen › Bestellung

5 Data analysis in molecular biology and evolution

Treffer erschließen

Aktive Filter
Datenquelle
- TIBKAT (entfernen)

Fach
Informatik
Mathematik
Technik
› mehr anzeigen

Medientyp
Buch
Report
Konferenzband
› mehr anzeigen

Format
Print
Elektronische Ressource
CD-ROM / DVD
› mehr anzeigen

Autor
Michalewicz, Zbigniew
Zgurovsky, Mikhail Z.
Albrecht, Alexander
› mehr anzeigen

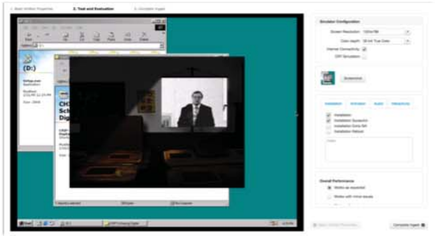
Erscheinungsjahr
2014
2013
2012
› mehr anzeigen

CD-ROM Collection Ingest Workflow

- Evaluate and prioritize data carrier migration
- License evaluation and rights clearance
- Creation of media images (e.g. ISO)

Workflow-Layer

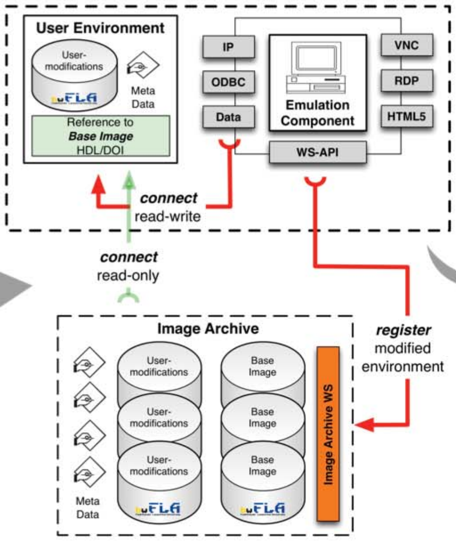
Ingest



Creation of tech. metadata

- Select and retrieve CD-ROM image
- Select an EaaS base image
- Install additional software (optional)
- Configure environment (optional)
- Evaluate object rendering
- Create and save a citable environment
- Output of technical metadata

Tailored Rendering Environment

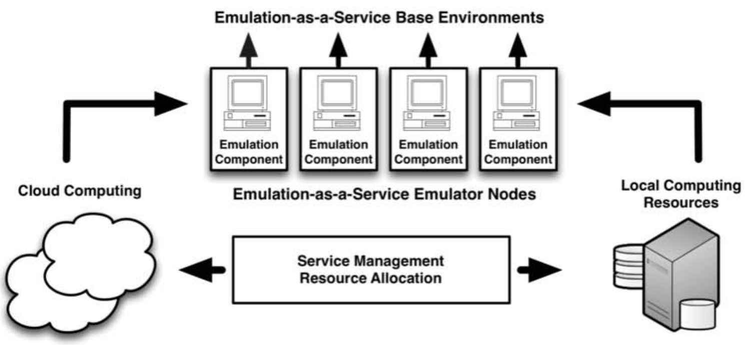


Access

Object rendering

- Load technical metadata
- Allocate computing resources
- Prepare emulator node
- Load and deploy environment
- Inject object (CD-ROM ISO)
- Re-enact environment
- Provide interactive access to users

Technical-Layer Emulation-as-a-Service



Demo

http://tib-test.bw-fla.uni-freiburg.de
user: tibtest
pw: tibtest2014

bwFLA Funktionale Langzeitarchivierung

The Digital POWRR Project: Enabling Collaborative Pragmatic Digital Preservation Approaches

Stacey Erdman
Northern Illinois University
97 Founders Memorial Library
DeKalb, IL 60115 USA
815-753-1004
serdman@niu.edu

ABSTRACT

The Digital POWRR Project has spent several years investigating scalable and practical digital preservation solutions that might be able to be implemented at smaller and less-resourced institutions.

General Terms

Communities

1. POSTER SUMMARY

The most well-known, robust digital preservation programs and initiatives have been initiated and shepherded at very large institutions, or at places with deep financial resources, often in collaboration with one another. Many less-well-resourced institutions create and hold digital objects worthy of preservation, but lack adequate resources to maintain them long-term. Working from the premise that digital preservation should be an attainable goal for everyone - even libraries and archives with restricted resources - the Digital POWRR project was conceived. Funded by a National Leadership Grant by the Institute for Museum and Library Services, Digital POWRR has been tasked with the in-depth investigation and testing of digital preservation solutions, tool and services, as well as positing potential business models and collaborative frameworks specifically designed for under-resourced libraries. Project activities are driven by five institutions of higher learning of varying sizes and resource levels drawn from across the state of Illinois: Northern Illinois University, Illinois State University, Western Illinois University, Chicago State University, and Illinois Wesleyan University. This poster will provide an encapsulation of our final white paper results, which will be published by the IMLS in late 2014, including a summary of our testing activities and results, details of collaborative endeavors within the project, an evaluation of the lessons learned and results gained, and a discussion of potential future directions for the project.

In addition to performing in-depth testing and evaluation of six robust digital preservation solutions (Archivematica, Curator's Workbench, Preservica, DuaCloud, and MetaArchive) POWRR partners spent considerable time developing and compiling a more comprehensive Tool Grid, in which smaller microservice-based tools or processes could be evaluated against the digital curation lifecycle. POWRR is presently aligning these research results and processes with other international initiatives, including seamlessly integrating our Tool Grid into the COPTR (Community Owned digital Preservation Tool Registry) tool registry produced by the SPRUCE project. We are also producing a series of day-long workshops to be held at various events across the United States throughout 2014, some in partnership with the National Digital Stewardship Alliance. By partnering with the NDSA, we hope to be able to increase the number of workshops that we are able to offer, and to be better positioned to reach our targeted audience of professionals working in lesser-resourced institutions of cultural heritage. The workshops will provide participants with a window into our testing activities, as well as give them an opportunity to perform some hands-on work with some simple microservices. The workshops will also detail the necessary steps in order to begin a publicity and education program for digital preservation activities within their own institution. We are also working with the lead counsel from the Educopia Foundation to create the necessary legal and technical foundational structure for institutions who are not part of an existing consortia to be able to form collaborations that enable distributed digital preservation endeavors. Documents produced from this work will be made available to the public, to help provide the necessary legal and organizational framework that organizations may need to get started with collaborative preservation initiatives.

2. ACKNOWLEDGMENTS

The author wishes to express sincere gratitude to the Institute of Museum and Library Services for funding the Digital POWRR project, as well as providing invaluable feedback and guidance along the way. The ambitious full scope of the project would not have been possible without numerous contributions from many staff members from our five partner institutions: Northern Illinois University, Illinois State University, Western Illinois Univeristy, Chicago State University, and Illinois Wesleyan University. Additionally, the project was guided by a expert Board of Advisors who provided ongoing feedback, critiques, and evaluation of all stages of our work. Board members include Amy Rudersdorf (Assistant Director for Content, Digital Public Library of America), Martin Halbert (Dean of Libraries and Associate Professor, University of North Texas), Steve Bromage (Executive Director, Maine Historical Society), Liz Bishoff (Principal Partner, The Bishoff Group LLC), Jerome McDonough (Associate Professor, University of Illinois at Urbana-Champaign), and Christopher Prom (Assistant University Archivist and Associate Professor, University of Illinois at Urbana-Champaign).

PRESERVING (DIGITAL) OBJECTS WITH RESTRICTED RESOURCES

POWRR

HTTP://DIGITALPOWRR.NIU.EDU

I. INTRODUCTION

Funded by a generous National Leadership Grant from the Institute of Museum and Library Services, the Digital POWRR Project has spent the last several years investigating scalable and practical digital preservation solutions that might be able to be implemented at smaller and less-resourced institutions. Digital POWRR has successfully leveraged the expertise and unique perspectives of cultural heritage professionals drawn from five institutions of higher learning of various sizes and resource levels from across the state of Illinois, including Northern Illinois University, Illinois State University, Western Illinois University, Chicago State University, and Illinois Wesleyan University. The Digital POWRR was tasked with the in-depth investigation and evaluation of scalable, sustainable digital preservation solutions for these smaller institutions. The team is also exploring potential business and implementation models for equitable access to digital preservation, so that these memory institutions can become a driving force in protecting their organizations' significant digital objects.

II. WHITE PAPER

Per the IMLS' orders, the Digital POWRR Project's main objective has been to produce a comprehensive white paper that summarizes all testing and evaluative activities undertaken for the grant. The POWRR team worked with the project's Board of Advisors to select a manageable number of digital preservation tools and services for the group to perform in-depth evaluations. These include Archivematica, Curator's Workbench, DuraCloud, Internet Archive, MetaArchive, and Preservica. It was agreed that testing should focus on the accessibility and usability of the tools/services by a practitioner at an under-resourced institution and the constraints they could possibly be working under, including:

- Outdated technical infrastructure
- Little to no budget for licensing fees, additional equipment, etc
- Little or no access to staff possessing technical skills (programmers, metadata or digital collections staff), or access to server administrators
- Limited personal technical skills
- "Lone Arranger" set-ups

Although each tool/service was evaluated against an OAIS-based digital preservation functionality matrix, written comments made by individual testers help ground our reviews in the practical world. Reviews also include pricing information when needed, and a description of the technical know-how it takes to get the service operational.

Testing was never viewed by the team as a way to "rank" tools and services in any way, but rather, as a way to help us provide some simple solution models for a community that may not have the time or energy to devote to sorting out exactly what all these things exactly do, and how exactly they relate to one another.

The white paper also includes case studies written by each contributing institution that provide an unvarnished view of the challenges and roadblocks that each one of us has been confronting when dealing with our digital collections. The white paper can be downloaded by visiting http://goo.gl/Mfqz6l (or scan the following QR code with your smartphone).

CURATOR'S WORKBENCH

archivematica.

INTERNET ARCHIVE

Meta-Archive COOPERATIVE

Preservica

III. COLLABORATION & DELIVERABLES

Over the course of the project, the team found a number of opportunities to leverage our research into a number of collaborative endeavors. These include:

1. Tool Grid/COPTR

In addition to the intensive testing conducted on the six major tools/services, POWRR team members also invested time into compiling a more comprehensive Tool Grid (http://digitalpowrr.niu.edu/toolgrid), in which smaller microservice-based tools or processes could be compared on the same evaluative matrix. The project was soon contacted by staff from the SPRUCE project to ask if we would be willing to collaborate on building a master digital preservation tool registry. This registry was later launched by the Aligning National Approaches to Digital Preservation initiative, and was named COPTR (Community Owned digital Preservation Tool Registry). The project team was delighted that our work could be incorporated into such a truly international endeavor, and happily supplied our tool evaluations for incorporation within COPTR, alongside other partners including the Digital Curation Centre, Digital Curation Exchange, the National Digital Stewardship Alliance, and the Open Planets Foundation.

COPTR

2. Educational Handouts

Early on in our research, team members realized that practitioners at smaller organizations may have better success in securing the resources for a digital preservation program if they had some basic assistance with advocacy. Working with a library communications consultant, our team was able to devise a set of one page educational handouts targeted at various stakeholder groups (Administration, IT, Content Creators, etc.) that provide an easy introduction to digital preservation and why it should particularly matter to them. These handouts are made available to the public under a Creative Commons license, and are provided in a modifiable format for practitioners to customize as needed.

3. (Duke) Data Accessioner

Duke Data Accessioner was developed at Duke University's David M. Rubenstein Rare Book and Manuscript Library as an easy way of migrating data off disks, while also enabling users the ability to run several microservices on the files as they are migrated to a more secure file server. The POWRR team collaborated with the software developer behind Duke Data Accessioner, Seth Shaw, to update and improve facets of the GUI, plug-in infrastructure, and metadata output.

4. Legal Frameworks

While testing the Collaborative MetaArchive Membership model, the team recognized that the development of appropriate legal relationships and organizational forms between collaborating institutions represented a significant barrier to smaller organizations seeking affordable methods of distributed digital preservation. The team decided to work with the legal counsel for the Educopia Institute to create the business models and legal framework for this type of partnership. All models and legal contracts created in this endeavor will be made publicly available under a Creative Commons license and provided in a modifiable format.

5. Internet Archive Tutorial

Several POWRR team members routinely work with small local historical societies, and are well-aware that these institutions frequently create digital objects with no thought of long term storage or preservation. We worked with a handful of these organizations to see how they might make best use of the free features of the Internet Archive. Team contacts were oftentimes volunteers with little computer sophistication and required additional instruction on how to use the service. POWRR created a basic tutorial to use with this audience, and has now made it freely available to the public.

IV. WORKSHOP

For some time, the POWRR team wrestled with the best mechanisms for disseminating our results to the widest possible audience. In consultation with the Board of Advisors, we decided that a traveling free or low-cost workshop would have the best potential for reaching (and helping!) our target audience. In planning our appearances, the team tried to aim for a fairly even distribution around the US. For some workshops, we elected to partner with some relevant conferences that archivists, librarians, digital humanists, and museum professionals would already be attending. For others, we partnered with the National Digital Stewardship Alliance and their membership base, or with other consortia-type arrangements.

Designed for practitioners with some previous understanding of digital preservation on a conceptual level, the workshop provides participants with a window into the nitty-gritty of our testing activities and into the tools and services that we have worked most closely with. Additionally, we give them an opportunity to perform some hands-on work with some simple microservices, through our demo and guided exercises with the Data Accessioner tool.

We created a small group exercise called the "3-3-3 Action Plan" to help participants take the necessary steps in order to begin a publicity and education program for digital preservation activities within their own institutions. It is our hope that these pragmatic, "hands-on" activities provide participants with the necessary confidence to return to their home institution and take the necessary first steps towards creating a digital preservation program.

3-3-3 Action Plan

V. LESSONS LEARNED

- Digital preservation is a small world...smaller than we originally thought.
- Don't fall into the trap of getting hung up on theory vs. practice. For some of us, it's time to embrace "good enough" digital preservation.
- The worry about "doing it right" has led to inaction and utter paralysis for many.
- Solutions are rapidly evolving...you likely will use different digital preservation tools and services throughout your career, so try to remain flexible!
- Additionally, you may find that you need to develop a toolbox approach to doing digital preservation - you may require a suite of tools, services, workflows, etc. rather than a turnkey approach.
- Bigger schools don't necessarily have it all figured out. Issues of larger scale and scope may make their priorities different than our own, but do not let this stop any fruitful collaborations from occurring.
- Smaller/less resourced institutions have unique advantages. We may have less administrative bureaucracy to navigate; less people to educate, or less restrictions on our computing environment - work these to your advantage!
- Vendors may change business models if you provide well-reasoned solid input, and can prove that your market segment might be better served at a different price point or service level.
- There are levels of triage that can be taken on your materials - the actions you take with them can be iterative. You needn't start with the implementation of a robust program.
- Planning and advocacy is an oftentimes neglected activity, but is vital to success with digital preservation. It is vital to know your organization's strengths and weaknesses, and to be able to strategize.
- Is is indeed possible to do hands-on technical exercises in a workshop full of people of varying skill levels...and it was very much what people want to experience in a workshop setting!
- When we submitted the grant, we were unsure what our reception would be, and if we would have a successful conclusion. It is gratifying to know that our initial hunches were all correct, and that there was a need in this area.
- We are glad that the project has truly has struck a chord with so many people, and has been able to provide assistance, advice, and inspiration to others.

QR CODES

In order to access the webpages cited within this poster via QR codes, you will need to use an app such as QR Scanner for iPhone or Barcode Reader for Android devices.

THE DIGITAL POWRR PROJECT:
ENABLING COLLABORATIVE PRAGMATIC DIGITAL PRESERVATION APPROACHES

VI. FUTURE PLANS

The POWRR project will draw to a close in November 2014, but team members hope that the project will live in through other endeavors. First, we are hoping that our upcoming workshops this year will allow us to identify smaller organizations that we can bring on-board and into the NDSA membership ranks. We will be seeking additional funding to continue our workshops, and will be revising and expanding the curriculum as we receive feedback from participants. We also want to take the workshop into a virtual environment. Financial accessibility is one of our biggest concerns, and this would remove the final hurdle of getting us to the people who need the information the most. Team members are also exploring possibilities for building the technical infrastructure for a collaborative digital preservation system in our region.

BY STACEY ERDMAN,
DIGITAL POWRR,
NORTHERN ILLINOIS UNIVERSITY

DESIGNED BY DANIEL M. KANEMOTO | www.exmortisfilms.com

E-ARK PROJECT – BEST PRACTICE SURVEY RESULTS ON ARCHIVING OF DIGITAL MATERIAL

Clive S G Billenness
University of Portsmouth
School of Creative Technologies
Portsmouth
+44 208 123 2782
clive.billenness@port.ac.uk

Kathrine H E Johansen
The Danish National Archives
Rigsdagsgaarden 9
DK-1218 Copenhagen K
+45 41 71 72 20
khej@sa.dk

David Anderson
University of Portsmouth
School of Creative Technologies
Portsmouth
+44 23 9284 5525
cdpa@btinternet.com

ABSTRACT

This paper describes the poster presented at iPres2014 by the EC-funded E-ARK Project, detailing the results of an international user survey recently conducted into Best Practice in the full life cycle of the long term preservation of digital records by archival organizations in the state and private sectors.

General Terms

Communities, Preservation Strategies, Preservation Workflows, Best Practice.

Keywords

Digital Archives, User Survey, E-ARK, EC, ICT-PSP, SIP, AIP, Pilot, e-infrastructure, digital archives, data mining, OAIS, Big Data, born-digital records, ingest, access

1. INTRODUCTION TO E-ARK

European **A**rchival **R**ecords and **K**nowledge **p**reservation (E-ARK) was launched in February 2014 and is a new, 3-year pilot project within the European Commission's ICT Policy Support Programme. With 16 partners in 11 EC countries comprising end users, research institutions and systems suppliers, its objective is to provide a single, scalable, robust approach capable of meeting the needs of diverse organisations, public and private, large and small, and able to support complex data types. E-ARK will demonstrate the potential benefits for public administrations, public agencies, public services, citizens and business by providing simple, efficient access to the workflows for the three main activities of an archive - acquiring, preserving and enabling re-use of information.

E-ARK will implement a number of pilot systems in different countries addressing challenges which differ in content and scale in order to create, by the end of the project, in 2017, a suite of openly-accessible end-to-end solutions capable of integration into third-party products and which will be sustained into the future.

Our work is worldwide the first attempt to bring together working elements of archival systems. As such it is an ambitious project which has several key features: creating standardized pre-ingest formats / specifications; expanding MoReq modules to be used as a key element of the infrastructure; using CMIS and Big Data techniques to promote new ways of access to digital archives, etc. It also addresses a wide range of users: public bodies, commercial institutions, individual citizens and researchers.

Our project will also provide a Digital Preservation Maturity Model which will enable organizations to not only assess their current performance, but also to measure improvement.

More information about the project is available from our website at www.eark-project.eu.

2. THE BEST PRACTICE SURVEY

2.1 Methodology

During the summer of 2014, the project conducted a survey into available best practices in Data Ingest, Available Requirements and Formats for AIP's and into the gaps between existing Access Solutions and users' needs.

We conducted an online survey to which responses were received from 184 respondents from 32 countries. Respondents covered a wide range of domains:

- Archives
- Private Companies
- Public Organisations
- Libraries
- Universities and
- Private Organisations

The questions within the survey were differentiated to be appropriate to each group.

We then followed up with person-to-person interviews with representatives from seven archives and four digital archiving solutions vendors.

The full results of our work are available for download from our website.

2.2 Available best practices in data ingest

We identified that two approaches must be differentiated: the ingest of whole systems and the ingest of individual records.

No widespread practices for records export can be identified. The most commonly used standards are ISO15489-1 and MoReq, but there seem to be lack of consensus on what is good practice for records export

Ingest workflows are generally considered to be in accordance with the OAIS model but the implementation varies greatly.

PIMAS steps can be used to describe/cover the most common steps in the ingest workflow.

The understanding of what constitutes a SIP varies greatly between organisations. Some consider simple computer folders as SIPs, others consider metadata standards as SIP and other again have defined a specific structure for SIPs.

Similarities between SIP formats can be found the use of metadata standards where METS, PREMIS and EAD emerge as significant.

Most SIP formats include the following four components:

- xml-file for describing the structure of the SIP,
- an xml-file for descriptive metadata,
- a unique identifier (UID), and
- a folder with content

2.3 Available requirements and formats for AIP’s

Based on a set of generic criteria for an ideal conceptual AIP, the following existing AIP formats were identified as the best existing AIP concepts:

- US Patent 13/219,630 Method And System For Preparing Digital Information For Long-Term Preservation
- BSI TR-03125 Preservation of Evidence of Cryptographically Signed Documents, Annex TRESOR-F: Formats and Protocols
- The AIP from Archivematica
- The AIP from ESSArch
- The German developed DA-NRW AIP format
- The AIP format from Preservica

For AIP containers, the three most significant formats are ACE (.ace), RAR (.rar) and Tape Archive (.tar), each of which was evaluated for its suitability for further use within the E-ARK Project. At this stage, Tape Archive is recommended as the most suitable format to be used in the ongoing E-ARK work.

2.4 Gap analysis between existing access solutions and users’ needs

Our study revealed that users are concerned with the following requirements for archival access services:

- Contemporary solutions that meet the standards of modern IT services
- Services that are easy to use i.e. that do not require specific technological or human skills
- Speed of access and usability and flexibility of services
- Integration or interoperability between different parts of an access service e.g. between finding aids and presentation tools
- Possibilities to search across Information Packages in data and metadata
- Functionalities that support their specific purpose

We concluded that:

- There is still limited experience with providing access to born-digital material
- Generally users’ needs are not met very well by existing services

The most prominent gaps are between existing solutions and users’ needs are:

- lack of flexible and modern solutions,
- lack of interoperability between components,
- lack of comprehensive metadata in finding aids,
- lack of functionalities to support use of data in presentation tools

3. NEXT STEPS IN E-ARK

Our project will now proceed to develop requirements and recommendations for the export of source records.

We will build an ingest workflow based on PIMAS.

We will include the four most common components of SIP’s in the E-ARK SIP format.

We will continue to develop our products with a persistent core but with sufficient flexibility to accommodate the specific needs of individual organizations.

4. TO FIND OUT MORE ABOUT E-ARK

A PDF version of our poster, together with copies of all our reports, can be downloaded from our project website:

www.eark-project.eu

At the website, it is also possible to sign up for a mailing list to be kept informed about developments in the project.

You can also follow our Twitter feed at @EARKProject.

E-ARK

E-ARK

European Archival Records and Knowledge Preservation

www.eark-project.eu

2014 Best Practice Survey On The Archiving Of Digital Material By Archival Institutions

WHAT WE DID

We surveyed 184 people in 32 countries representing:

- Archives
- Private Companies
- Public Organisations
- Libraries
- Universities
- Private Organisations
- 'Others'

HOW WE DID IT

We used a combination of on-line questionnaires with questions appropriate to each group and

Follow-up person-to-person interviews with

- 7 Archives
- 4 Digital Archiving Solutions Vendors

WHAT WE DISCOVERED

AVAILABLE BEST PRACTICES IN DATA INGEST

Ingest: whole systems vs. single records?
Follow standards/guidelines for records export?

- No widespread practices for records export can be identified. The most commonly used standards are ISO15489-1 and MoReq but there seem to be lack of consensus on what is good practice for records export
- Ingest workflows are generally considered to be in accordance with the OAIS model but the implementation varies greatly. PIMAS steps can be used to describe/cover the most common steps in the ingest workflow.
- The understanding of what constitutes a SIP varies greatly between organisations. Some consider simple computer folders as SIP, others consider metadata standards as SIP and other again have defined a specific structure for SIPs
- Similarities between SIP formats can be found the use of metadata standards where METS, PREMIS and EAD emerge as significant.
- Most SIP formats include the following four components:
  - ⇨ xml-file for describing the structure of the SIP,
  - ⇨ an xml-file for descriptive metadata,
  - ⇨ a unique identifier (UID), and
  - ⇨ a folder with content

AVAILABLE REQUIREMENTS AND FORMATS FOR AIP's

Storage: Does the organisation have an AIP format?
Different AIPs for different content types?

Based on a set of generic criteria for an ideal conceptual AIP, the following existing AIP formats are identified as the best existing AIP concepts:

- US Patent 13/219,630 Method And System For Preparing Digital Information For Long-Term Preservation
- BSI TR-03125 Preservation of Evidence of Cryptographically Signed Documents, Annex TRESOR-F: Formats and Protocols
- The AIP from Archivematica
- The AIP from ESSArch
- The German developed DA-NRW AIP format
- The AIP format from Preservica

For AIP containers the report identifies the three most significant formats to be ACE (.ace), RAR (.rar) and Tape Archive (.tar) and subsequently evaluates their suitability for further use within the E-ARK Project. Tape Archive is recommended as the most suitable format to be used in the onward E-ARK work.

GAP ANALYSIS BETWEEN EXISTING ACCESS SOLUTIONS AND USERS' NEEDS

Does the organisation provide access to digital material?
Which content types?
What platform(s) are used to provide access?

The study revealed that users are concerned with the following in archival access services:

- Contemporary solutions that meet the standards of modern IT services
- Services that are easy to use i.e. that do not require specific technological or human skills
- Speed of access and usability and flexibility of services
- Integration or interoperability between different parts of an access service e.g. between finding aids and presentation tools
- Possibilities to search across Information Packages in data and metadata
- Functionalities that support of their specific purpose of use

The report concludes that:

- There is still limited experience with providing access to born-digital material
- Generally users' needs are not met very well by existing services

The most prominent gaps are between existing solutions and users' needs are:

- lack of flexible and modern solutions,
- lack of interoperability between components,
- lack of comprehensive metadata in finding aids,
- lack of functionalities to support use of data in presentation tools

✔ E-ARK will now develop requirements / recommendations for records export
✔ We will build an ingest workflow based on PIMAS
✔ We will include the 4 most common components of SIPs in the E-ARK format
✔ E-ARK Products will be flexible but with persistent core.

To download the full reports or find out more about our project, please visit our website

University of Portsmouth
REPUBLIKA SLOVENIJA MINISTRSTVO ZA KULTURO ARHIV REPUBLIKE SLOVENIJE
DLM FORUM FOUNDATION
DigitalPreservationCoalition
AIT AUSTRIAN INSTITUTE OF TECHNOLOGY
THE NATIONAL ARCHIVES OF NORWAY
STATENS ARKIVER The Danish National Archives
GOBIERNO DE ESPAÑA MINISTERIO DE HACIENDA Y ADMINISTRACIONES PÚBLICAS
RAHVUSARHIIV THE NATIONAL ARCHIVES OF ESTONIA
archívum
MAGENTA aps
KEEPSOLUTIONS University of Minho SPIN-OFF
ES Solutions
TÉCNICO LISBOA
ama

THE E-ARK PROJECT IS CO-FUNDED BY THE EUROPEAN COMMISSION
UNDER THE ICT-PSP PROGRAMME

ICT PSP

State Records NSW Digital Archives Poster

Richard Lehane
State Records NSW
PO Box 516
Kingswood NSW 2747
+61 02 9673 1788
richard.lehane@records.nsw.gov.au

Danny Archer
State Records NSW
PO Box 516
Kingswood NSW 2747
+61 02 9673 1788
danny.archer@records.nsw.gov.au

Cassie Findlay
State Records NSW
PO Box 516
Kingswood NSW 2747
+61 02 9673 1788
cassandra.findlay@records.nsw.gov.au

ABSTRACT

In this paper, we describe a poster on the State Records New South Wales approach to digital archives.

General Terms

infrastructure, strategic environment, preservation strategies and workflows, theory of digital preservation, case studies and best practice

Keywords

recordkeeping, archives, State Records New South Wales

1. INTRODUCTION

The business of government is carried out using a wide variety of tools and technologies. Requirements for records of that business to be retained permanently as State archives are technology agnostic, and can apply to virtually any format or system type. In addition, records are, by definition, heavily dependent on context and relationships. Decoupled from context they lose their evidential value. These factors all pose major challenges when, as the State's archive, it is necessary to take on the responsibility, under law, for the accessibility, integrity and usability of these records - forever.

The State Records approach to the capture and retention of systems of born digital records into the archive is an innovative solution – or set of solutions – to the problem of the preservation of digital records. The State Records New South Wales Digital Archives team poster explains the processes, tools and technologies that support the team's migration project based approach to the capture of records into the archive. It will also describe the tools and frameworks that the team is using for the continued management of the records and their contextual information over time.

2. ABOUT THE POSTER

The State Records NSW digital archives poster will explain:

- the phases of State Records' Digital Archives Migration Methodology
- tools, frameworks and services adopted by the team to address preservation, context capture and searching challenges; and
- applications developed by the team to execute migration and search processes across disparate sets of digital records and their metadata.

2.1 A structured approach

The Digital Archives Migration Methodology provides a structured framework. Each migration is based on a full understanding of the requirements of the digital records informing a tailored migration workflow that is planned and accountable.

2.2 Tools and technologies

Tools and services adopted by the team to solve particular preservation problems have been drawn from initiatives at the national and international level including the National Archives of the UK's PRONOM project and projects at California Digital Libraries as well as open source tools such as DROID, Exiftool, Apache Tika and more.

The applications developed in house by State Records and released progressively as open source via GitHub are described in this poster include:

- the Migrate tool for executing preservation and metadata rules and moving records and their metadata into the Digital Archives repository;
- the Metadata Registry allowing State Records to gradually build a schema of terms to which metadata from new projects can be mapped or added;
- Preservation Pathways, for recording and sharing decisions made about file format conversions, linked to the PRONOM technical registry but able to operate independently of it; and
- a search application for records and metadata.

This poster and accompanying presentation the State Records Digital Archives team share an innovative but pragmatic approach taken to the challenge of meta-recordkeeping; the capture, preservation, accessibility and management of myriad recordkeeping systems in one environment.

For more information go to:
http://www.records.nsw.gov.au/digitalarchives

Migration Methodology

State Records
Digital Archives

The Digital Archives Migration Methodology supports the transfer of digital records from NSW Government agencies to the Digital State Archive. Rather than adopting a single approach for all such transfers, State Records NSW defines custom migration plans to suit the particular requirements of different sets of records. Each transfer is managed as a separate project. The methodology is a framework to guide these projects.

By blending project management and data migration techniques, the Digital Archives Migration Methodology provides a structured and planned approach to each migration project. It also permits flexibility, for dealing with many types of migrations, from very simple ones to complex ones involving many record types and stakeholders.

1. Project Planning Phase

The Project Planning phase establishes a framework for the migration project.

The purpose of this phase is to define the project goals and identify stakeholders, risks, and resources. The depth of planning required in this phase will vary depending on the complexity of the project. For example, a project involving the transfer of a single audio file might be very small in comparison with a project involving the transfer of a business or email system.

The key deliverable of this phase is the project plan.

Tools

Basecamp *Adopted*
https://basecamp.com
Basecamp is a web-based project management tool. It is used to collaborate with project participants, plan and schedule actions, and share documentation.

Digital Archives Migration Methodology *Built*

Key Relationships

State Records Act 1998
S. 29 of the State Records Act allows State Records to issue guidelines about how records are to be made available to it. The guidelines apply to records in any format, including digital records.

2. Migration Planning Phase

The goal of the Migration Planning phase is to develop the migration plan.

The migration plan is a document that identifies and documents the activities to be carried out during the migration of a recordkeeping system into the Digital State Archive. These activities include file format migration, metadata mapping, and data transformation.

Like project plans, migration plans are tailored to suit the requirements of the particular project. But this doesn't mean re-inventing the wheel each time: the decisions and lessons learned in each project are documented and can be re-used in subsequent projects.

Tools

DROID/PRONOM *Adopted*
http://apps.nationalarchives.gov.uk/PRONOM/Default.aspx
The National Archives UK's technical registry and file format identification tool are used for canonical identification of file formats.

Apache Tika *Adopted*
http://tika.apache.org/

ExifTool *Adopted*
http://www.sno.phy.queensu.ca/~phil/exiftool/
Metadata extraction tools such as Tika and ExifTool are used to supplement agency-provided metadata.

Preservation Pathways Registry *Built*
http://www.records.nsw.gov.au/digitalarchives/pathways/
When digital records are migrated to the digital archives, file formats are assessed for their longevity and accessibility. In some cases a transformation is recommended. Transformations are registered in Preservation Pathways with information about the input and target file formats (using IDs from the National Archives UK's PRONOM registry as well as information about the tool or process used to perform the transformation). If file format IDs are not available from the PRONOM registry, then temporary State Records NSW IDs are registered pending the creation of a PRONOM ID.

The Preservation Pathways Registry is a Java web application.

Metadata Registry *Built*
http://www.records.nsw.gov.au/digitalarchives/metadata/
The Digital Archives Metadata Registry allows Digital Archives staff to progressively register preferences for published metadata terms (e.g. Dublin Core) to represent common metadata elements in the digital archives. It also allows Digital Archives staff to progressively coin new terms (by providing a URI and description) to represent metadata elements in the digital archives for which no suitable published term can be identified.

The Metadata Registry is implemented with JSON Schema and using Github.

Key Relationships

Disposal authorities
Part 3 of the State Records Act 1998 prohibits the disposal of State records except where it is authorised. Under the Act, State Records can give permission for disposal. The usual means by which State Records permits disposal is through the approval of retention and disposal authorities.

Only digital records required as State archives under an authorised disposal authority may be transferred to the Digital State Archive.

Access Directions
Part 6 of the State Records Act creates a framework for regulating public access to State records which have been in existence for at least 30 years (the 'open access period'). Public offices are required to make an access direction (to determine whether the records are open or closed to public access) for all their records which are in the open access period.

Agencies transferring records to the Digital State Archive must ensure that those records are covered by current access directions.

Series control system
State Records implements the Australian Series System to describe and control the State Archives. Digital archives are linked to this system by links to agencies and series.

3. Migration Phase

During Migration phase, the migration plan is executed. It is in this stage that any necessary preservation activities are performed.

Tools

Aspose *Adopted*
http://www.aspose.com/
A set of file format APIs for Java. Used for transformation of a number of common file formats.

EMC Isilon *Adopted*
http://www.emc.com/domains/isilon/index.htm
EMC Isilon scale-out network-attached storage is used to store digital archives. Advantages of this system include scalability, integrity, and automated replication.

Pairtree *Adopted*
https://wiki.ucop.edu/display/Curation/PairTree
A modified version of California Digital Library's Pairtrees for Object Storage protocol is used to manage the storage of digital objects. Pairtree is a filesystem hierarchy.

MongoDB *Adopted*
http://www.mongodb.org/
Metadata is stored as JSON documents in the filesystem and copied into a MongoDB instance for search and reporting

Apache Solr *Adopted*
http://lucene.apache.org/solr/
Full-text is extracted wherever possible. This full-text is stored in a Solr instance for search.

Migrate tool *Built*
A workflow tool that assigns unique identifiers to digital objects, validates consignments, and moves records into the Digital State Archive.

The Migrate tool is a Java command-line application.

Key Relationships

Search
http://search.records.nsw.gov.au/search
The Digital Archives team at State Records is responsible for Search, the main finding aid for the State Archives collection. Contents of the Digital State Archive are discoverable via Search.

4. Project Closure Phase

The Project Closure phase closes the project and identifies any required post-project activities (such as the disposal of source records).

Tools

Basecamp *Adopted*
http://apps.nationalarchives.gov.uk/PRONOM/Default.aspx
Basecamp is a web-based project management tool. It is used to collaborate with project participants, plan and schedule actions, and share documentation.

Receipt *Built*
http://apps.nationalarchives.gov.uk/PRONOM/Default.aspx
This tool provides agencies with a human and machine readable manifest of digital records successfully migrated into the Digital State Archive.

The Receipt tool is a Java command-line application.

Key Relationships

GA33 Source Records that have been migrated
This is a disposal authority that provides for the authorised disposal of State Records that have been used as the source records for successful migration projects. It is under this authority that public offices are permitted to destroy source records post-transfer to the Digital State Archive. This authority mandates a minimum six months retention period for quality assurance purposes.

Lessons learned in developing digital preservation tools the right way (and the wrong way)

Paul Wheatley
Paul Wheatley Consulting Limited
Leeds
West Yorkshire
@prwheatley
paulrobertwheatley@gmail.com

ABSTRACT

The digital preservation community has had a chequered history in developing software tools to perform operations essential for preserving digital data. Poor technology choices, half measures in adopting open source approaches, insufficient engagement with users, finite project funding and an array of other challenges have hampered tool development. Gaps in capability are common but even where tools have been created to take on a particular problem, they often face patchy support and an uncertain future. Lessons have however been learned from mistakes that have been made in the past. User engagement and an agile development approach can focus solutions on real problems. Adoption and expansion of existing tools (sometimes from outside of this community) can yield greater and more dependable results. Focused designs can make adoption easier. Sharing the experimentation and data behind tool development and assessment can be as invaluable as the tools themselves. This paper provides an outline of lessons learned from developing digital preservation tools across JISC and EU funded digital preservation projects, such as PLANETS and SCAPE and more recently from agile hackathon and mashup events run by SPRUCE.

Keywords

Digital Preservation, User Requirements, Digital Preservation Tools, Open source development

1. WHY CAN'T WE HAVE TOOLS THAT JUST WORK?

At iPRES2012 Steve Knight noted dissatisfaction in our preservation tools in a review of the previous decade of digital preservation: “Tools like DROID and PRONOM etc. didn’t work properly then, and they still don’t work properly now. The wish list from this year’s Future Perfect Conference (New Zealand) did not differ that much from the wish list four or ten years ago.” [1] Even tools that have (at least previously) seen considerable use within the community, such as JHOVE, are facing an uncertain future [2]. There are an array of reasons behind this:

- Many community created tools have struggled to survive once grant funding ended
- Tools from outside of the digital preservation community have often been overlooked
- Users have had insufficient say in the focus and design of preservation tools resulting in a mismatch between genuine user needs and preservation capabilities
- Organisations have in theory adopted open source development approaches, but this has often gone no further than depositing software in a code repository at the end of a project
- Little thought has been given to how new tools will be integrated with existing workflows
- A lack of even basic testing has resulted in a painful installation experience for many users

As Johan van der Knijff put it: “Why can't we have digital preservation tools that just work?” [3]

Lessons have been learned from the last 15 years of digital preservation research and development. Most of the unwieldy applications developed by the PLANETS Project [4] have not seen significant uptake, but focused preservation tools such as Jpylyzer [5], developed by SCAPE [6], have already been adopted by various organisations and have been embedded in workflow tools such as Goobi [7]. Engagement with existing tools that offer much to the preservation community, such as Apache Preflight [8], has yielded significant improvements to the tool. Pairing developers with practitioners in SPRUCE Mashups [9] has highlighted the rapid progress that can be made in solving preservation problems in even a short space of time (sometimes resulting in new preservation tools). These new approaches represent a sea change in the community which is beginning to focus more on practical experimentation, to share and exchange ideas with others using social media and to engage more readily with existing open source projects.

2. LESSONS LEARNED IN DEVELOPING PRESERVATION TOOLS

The poster will present the following lessons learned in developing preservation tools, and is adapted significantly from the SPRUCE Mashup Manifesto [10]. This in turn is based upon experiences in exploring, and in some cases solving, over 150 specific preservation challenges [11].

- **Be agile in your first steps**
  - Develop/prototype in short bursts, then demo and get feedback from your practitioner/user
  - If you don’t achieve results within a few hours, you are probably doing it wrong. Try a different approach
  - Get crude results quickly, perfect and polish later
  - Scripting languages can be useful for delivering quick results
- **Re-use, don't re-invent the wheel**
  - Most problems have already been solved, although often not by the preservation community

- Experiment with existing solutions first
- Someone else will have experience of other tools to try. Twitter can make the connections. Check the COPTR tools registry [12]
- Re-use existing code before writing any of your own. Existing code comes with existing users (who test and report bugs) and existing contributors. Exploit this where possible!

- **Keep it small, keep it simple**
  - Functional preservation tools should be atomic
  - Modularise in the face of growing requirements
  - Think about how someone else will integrate your tool in a workflow. Make it easy for Preservica, Rosetta, Archivematica and the rest to incorporate your code
- **Make it easy to use, build on, re-purpose and ultimately, maintain**
  - Test driven development simplifies subsequent maintenance
  - Share your source
  - Automate your build
  - Package for easy install
- **Share outputs, exchange knowledge, learn from each other**
  - Write up your experiences and share them (sharing less than successful experiences is just as valuable as successful ones!)
  - Publish the data you generate. This tool->this data->these results
  - Shout about it, blog it, tweet it, and add a tool registry entry to COPTR

Boxouts and examples will be used to expand on key points from the above.

3. REFERENCES

[1] Steve Knight quote in: Angevarre, Inge, NCDD Blog. http://www.ncdd.nl/blog/?p=3338

[2] Gary McGath's blog. http://fileformats.wordpress.com/2014/03/08/statejhove/

[3] Johan van der Knijff's blog. http://www.openplanetsfoundation.org/blogs/2014-01-31-why-cant-we-have-digital-preservation-tools-just-work

[4] Planets Project website. http://www.planets-project.eu/

[5] Jpylyzer, JP2 validation tool. http://openplanets.github.io/jpylyzer/

[6] Scape Project. http://www.scape-project.eu/

[7] Goobi, digitisation workflow management tool. http://www.digiverso.com/en/products/goobi

[8] Preflight, PDF validation library, http://pdfbox.apache.org/cookbook/pdfavalidation.html

[9] Paul Wheatley and Maureen Pennock. Supporting practical preservation work and making it sustainable with SPRUCE. *iPRES 2013 proceedings*. http://purl.pt/24107/1/iPres2013_PDF/Supporting%20practical%20preservation%20work%20and%20making%20it%20sustainable%20with%20SPRUCE.pdf

[10] SPRUCE Mashup Manifesto. http://wiki.opf-labs.org/display/SPR/The+SPRUCE+Mashup+Manifesto

[11] Digital Preservation and Data Curation Requirements and Solutions, http://wiki.opf-labs.org/display/REQ/Digital+Preservation+and+Data+Curation+Requirements+and+Solutions

[12] Community Owned digital Preservation Tool Registry (COPTR), http://coptr.digipres.org/

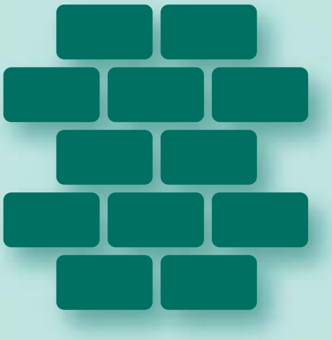
Lessons learned in developing digital preservation tools the right way (and the wrong way)

You know that something has gone wrong when the data outlast the preservation solutions designed to preserve them. Short term project funding has kicked off many digital preservation developments, but it hasn't often led to strong or sustainable results. New developments need to fit well within the context of existing infrastructure and solutions, they need to have a roadmap, some sensible governance and a maintenance plan that is realistic given an uncertain funding outlook. Most importantly, a community of users and digital preservation expertise needs to be at the heart of this approach. This is vital to ensure that the right developments are made, using sensible technologies and in a way that others can maintain on into the future.

Engage with the community

Engaging with the community from the very beginning enables new work to be shaped to the needs of real users. With the users in the driving seat, and the community coming along for the journey, the support is there to ensure results are of a high quality.

Key lessons learned:
* Capture and *share* the requirements
* Consult with the community before you start and work to solve their problems
* Work with actual examples or user data
* Designate a problem owner and a solution provider
* Facilitate frequent engagement between them

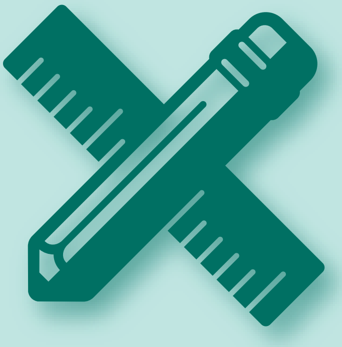


Build on existing work

There are many examples of digital preservation development that has gone it's own way, despite existing solutions from within or outside of the community. This duplication is incredibly wasteful. Engaging with existing work, even if it's outside of our community, *can* have a real impact and and reduce effort.

Key lessons learned:
* Thorough literature review should result in action!
* The outside world will care if we engage
* Support for missing digital preservation requirements *can* be added to existing work

Design for longevity



Digital preservation developments should of course employ digital preservation principles. New work should be designed for purpose and designed to last. Good management, planning, and/or software development techniques should be employed to ensure quality and sustainability.

Key lessons learned:
* Independent review can catch many issues
* Choose a technology and/or medium that is sustainable
Make preservation tools:
* Focused and atomic so they can be integrated easily
* That embody genuine open source techniques and tools
* Easy to test, enhance and maintain with community effort

Ally with a custodian

With so many new developments relying on external funding, the creator is often a project with a short lifespan. With a more long lived organisation engaged as a custodian from the very beginning, developments can benefit from their experience, community, support skills and stewardship. This might include the Open Planets Foundation, the Digital Preservation Coalition or the National Digital Stewardship Alliance.

Key lessons learned:
* Involve the custodian from the beginning of the development
* Draw on their skills, and community to ensure the development meets real needs
* Choose technologies that reduce barriers to interaction with users, and simplify the maintenance of resources
* For web hosted results, seek locations that will survive beyond the life of a project
* Remove legal barriers to resuse by ensuring all your results are published with clear licenses
* Separate (and hide) uninteresting operational details from real results on retired project websites. Even better, transfer the best results to a longer lasting home

Image credits:
"Design" by Scott Lewis, "Building Restoration" by Sean Connolly, "Institution" designed by James Fenton, and "Community" by Edward Boatman. All are from TheNounProject.com

Paul Wheatley Consulting Ltd
Digital Preservation Services
http://bit.ly/paulrobertwheatley

Preservation of Web Content – An Emulation-based Case Study

Dennis Wehrle, Thomas Liebetraut and Klaus Rechert
Albert-Ludwigs University Freiburg
Hermann-Herder Str. 10
79104 Freiburg i. B., Germany
firstname.lastname@rz.uni-freiburg.de

ABSTRACT

A significant amount of cultural activity has moved to the world wide web preservation of web content is becoming increasingly important. But with growing complexity and dynamics of web sites, a harvesting approach to web preservation has its limitations. We present a web preservation case study using a functional approach both for preservation and presentation. By using Emulation-as-a-Service (EaaS), a complete web site (web server, content management system, database) can be preserved and functionally re-enacted efficiently on demand.

General Terms

Case Studies and Best Practice

Keywords

Emulation, Web Preservation

1. INTRODUCTION

Preservation of electronic publications, especially preservation of web content is becoming a crucial task for memory institutions since a significant amount of cultural activity has moved to the world wide web. Hence, to reflect and preserve a significant part of modern cultural history, preservation of web content is indispensable.

By design, the world wide web is an open medium developed to ease access to information and to enable everyone to publish in a convenient and cost-effective way. Such a highly volatile medium, however, is problematic from a long-term preservation perspective. Web publications are neither centrally coordinated nor self-contained. A short technical life-cycle and fast changing technological trends add further (technical) complexity both to the acquisition phase and for securing long-term access.

In general, there are two technical ways to acquire web content, either from a consumer's or from a producer's perspective [1]. The consumer's approach is to use a web crawler that systematically "harvests" web pages in a similar way a normal user is surfing the web. This way, a broad range of content covering various domains can be collected with quite simple technical tools and infrastructure. However, by harvesting web sites, neither completeness nor consistency of acquired content can be guaranteed [2]. In case of complex, highly dynamic sites or if completeness or consistency matter, a second option is to preserve a web site's content directly at the producer's site, i.e. preserving the complete web server.

We present a web preservation case study using a functional approach both for preservation and presentation. By using Emulation-as-a-Service (EaaS), a complete web site (web server, content management system, database) can be preserved and functionally re-enacted efficiently on demand. Furthermore, several environments of that time, e.g. web browsers and various multimedia plug-ins, are available to access the preserved site, hence providing a feasible option for authentic presentation.

2. A WEB PRESERVATION CASE STUDY

Preservation of dynamic sites by harvesting its content using a traditional web crawler can be challenging. As a result, we have been approached by an e-learning web site owner who provides an interesting use-case for functional web site preservation. The web site contains both static pages as well as pages generated dynamically.

A functional web archiving approach first requires an analysis and preparation step, ideally carried out while the machine is still operational. In a second step, the physical machine is to be migrated to a virtualized or emulated environment. Finally, the preserved machine is prepared for on-demand re-enactment, providing functional access to its content.

The web server owner had already performed the imaging process and provided us with a 66 GByte disk image. The web server runs a SUSE Linux Enterprise Server 9[1] 64-bit operating system originally installed in early 2006. For content management, a ZOPE instance (Version 2.9.8) on top of a MySQL database is used. To prepare the machine for running in a virtualized / emulated environment, some minor issues had to be solved first, such as changing the device name of the root partition from originally *hda* to *sda*. For our use-case we also conducted some experiments on migrating a virtualized system (VirtualBox) to an emulated one (QEMU). After the migration, at the first boot the system complained about a changed graphics card, but after

[1] Linux sf4200-88 2.6.5-7.252-smp x86_64

accepting to reconfigure the system automatically the machine was fully operational.

For this scenario, EaaS emulation components provide the basis for preservation and replication of individual systems, e.g. the web server or auxiliary database servers. As these systems were not independent from each other, representing them as individually preserved computers is not sufficient. Instead, they usually require a specific network configuration including dedicated hostnames and routes to each other or may even require further hosts to be available.

Consequently, individual emulation components need to be interconnected via a dedicated virtual network infrastructure, replicating original network infrastructure and thereby fulfilling the individual machine's technical *expectations* on their working environment. As we had little extra information about the machine's original environment, in a second step implicit and explicit dependencies had to be determined manually. One such dependency was the network interface expecting a fixed public IP address and hostname together with a fixed gateway. These expectations on the external environment have been modeled as a virtual network environment, making it possible to reach the web site within the virtual network via its original public IP address.

Efficient Replication on Demand

To re-enact the web site's network structure including all its auxiliary resources within the EaaS framework, a technical description of all the network components and their configuration is required. Individual computer systems are configured within the EaaS framework using the *emulation environment* meta-data. This technical description of a computer platform contains information about hardware required to replicate the environment like processor type and hard disk drives and their configuration as emulated nodes.

To describe a complete network structure and the connection between individual environments, a *network environment* description is required. It constitutes the configuration meta-data necessary to rebuild a virtual network and describes all network components required to replicate the original system setup. This network environment is divided into individual subnetworks that correspond to the subnetworks found in IP-based networks and thus defines the global network structure. In order to access the web site by its original hostname a local DNS resolver was added to the network.

Usually, a network node represents a preserved computer environment, e.g. the web server or client platforms. Once they are preserved and defined using an emulation environment, they can be referenced in the node definition using a persistent identifier (currently HDL). The EaaS framework is then able to re-enact the network environment using these previously preserved computer environments, hence authentically reproducing the original environment.

In a web preservation scenario, however, it may not be sufficient to preserve just the server side but to also provide legacy clients with web browsers and operating systems of the web content's era. Once such a client system is available as an emulation environment within an EaaS framework, it can be incorporated into the network environment just like any other node. Henceforth it is possible to access the web site (provided by a preserved web server) in a functional way, either with an emulated client, providing a web browser of the web site's creation time, or using a current computer system (Fig. 1).

As a last step, several clients have been prepared and can be used with the web-server setup.

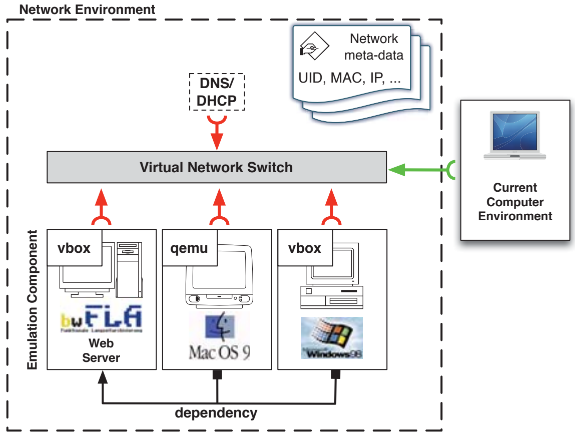


Figure 1: Web server and functional access options

While the technical network environment description allows the EaaS framework to replicate a preserved web site's network infrastructure, individual network nodes may depend on others. For instance, client nodes designed to access a web site may require this web site to be available before the user is able to access it or a content management system may require a database server to be available before it is instantiated. For this, each network node in a virtual network environment can declare service dependencies. The EaaS framework can use this information to determine which emulation component has to be started first and wait for it to be completely started. This way, it is possible not only to preserve a web server instance including its dependencies like database or content delivery servers, but also to re-enact a complex setup in a single step. Furthermore, a multitude of client template nodes can be provided with different operating systems running various web browsers, each instantiated independently just depending on the user's preferences.

3. CONCLUSION

By using a functional approach for web preservation it is possible both to preserve a specific web instance completely with additional services and dependencies, as well as provide authentic user-access by using client technology of the web site's time. Furthermore, the EaaS framework allows an efficient on-demand re-enactment with flexible resource allocation by providing a comprehensive set of technical metadata describing both individual machines and their original (networked) environment as well as their inter-dependencies.

4. REFERENCES

[1] J. Masanès. *Web Archiving*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2006.

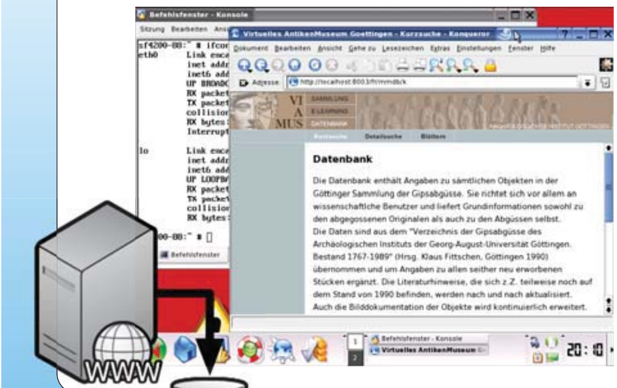
[2] Z. Xie, H. Van de Sompel, J. Liu, J. van Reenen, and R. Jordan. Archiving the relaxed consistency web. In *Proceedings of the 22nd ACM international conference on Conference on information and knowledge management*, CIKM '13, pages 2119–2128. ACM, 2013.

Preservation of Web Content
An Emulation-based Case Study

Dennis Wehrle, Thomas Liebetraut and Klaus Rechert

University of Freiburg, Germany

Original Environment (Server)


Befehlsfenster - Konsole
Virtuelles AntikenMuseum Goettingen - Kurzsuche - Konqueror
Dokument Bearbeiten Ansicht Gehe zu Lesezeichen Extras Einstellungen Fenster Hilfe
Adresse
http://localhost:8003/fr/mmdb/k
VIAMUS
SAMMLUNG
E-LEARNING
DATENBANK
Kurzsuche
Detailsuche
Blättern
Befehlsfenster - Konsole
Virtuelles AntikenMuseum
20:10
WWW

Datenbank

Die Datenbank enthält Angaben zu sämtlichen Objekten in der Göttinger Sammlung der Gipsabgüsse. Sie richtet sich vor allem an wissenschaftliche Benutzer und liefert Grundinformationen sowohl zu den abgegossenen Originalen als auch zu den Abgüssen selbst.
Die Daten sind aus dem "Verzeichnis der Gipsabgüsse des Archäologischen Instituts der Georg-August-Universität Göttingen. Bestand 1767-1989" (Hrsg. Klaus Fittschen, Göttingen 1990) übernommen und um Angaben zu allen seither neu erworbenen Stücken ergänzt. Die Literaturhinweise, die sich z.Z. teilweise noch auf dem Stand von 1990 befinden, werden nach und nach aktualisiert.
Auch die Bilddokumentation der Objekte wird kontinuierlich erweitert.

Technical Metadata

```
<networkEnvironment
xmlns="http://bwfla.bwl.de/common/datatypes">
<id>9997</id>
<description><title>Viamus Demo</title></description>
<network>

<!-- client -->
<emulatorNode>
  <hwaddress>10:23:45:67:89:10</hwaddress>
  <emulationEnvironmentRef>
    <id>9999</id>
  <emulationEnvironmentRef>
</emulatorNode>

<!-- server -->
<emulatorNode>
  <emulationEnvironment>
    <id>9898</id>
    <description>
      <title>Viamus Web Server</title>
    </description>
    <arch>x86_64</arch>
    <drive>
      <data>binding://main_hdd</data>
      ...
    </drive>

    <binding id="main_hdd">
      <url>hdl:11270/...</url>
      <access>cow</access>
    </binding>

    <nic>
      <hwaddress>00:14:4f:0f:73:94</hwaddress>
    </nic>
  </emulationEnvironment>
</emulatorNode>

</network></networkEnvironment>
```

Workflow

1.) Analysis & Preparation

While the machine is still operational, an assessment of hardware, software and configuration (e.g. operating system, memory, network configuration, ...) is required.

2.) Migrate physical machine to an emulated environment

Every system is bound to a specific configuration of its technical environment. Migrating the system into an emulated or virtualized environment is the first step towards a stable environment, as this procedure exchanges the specific hardware configuration with a well documented and well understood configuration (e.g. standard network card).

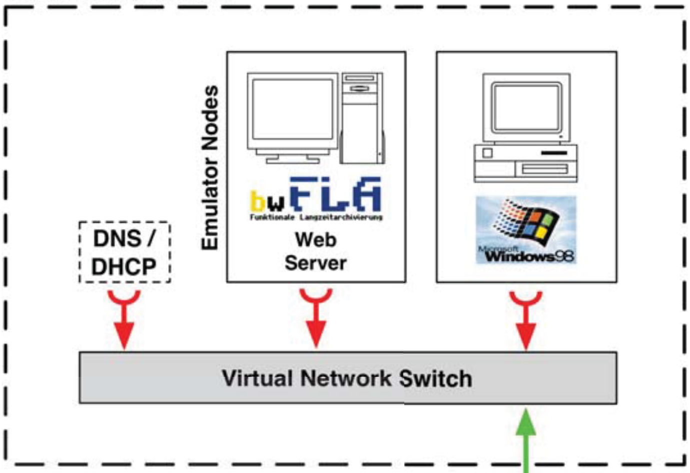
3.) Determine & resolve dependencies

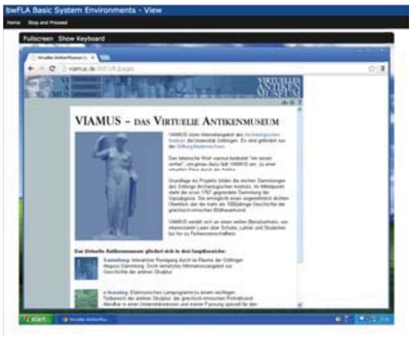
To enable long-term preservation all explicit as well as all implicit dependencies and their functional expectations, i.e. connection to data storage or external data base, have to be identified. Ideally all dependencies should be dealt with to become independent of future changes.

4.) Generate technical metadata

An EaaS instance should be replicable deterministically, i.e. run the same configuration later without configuring all components again. Hence a comprehensive description of an emulated environment is needed. This allows for exchanging EaaS components without losing existing environments and to outlive technical life-cycles.

Network Environment




bwFLA Basic System Environments - View
Home
Stop and Proceed
Fullscreen
Show Keyboard
viamus.de
VIRTUELLES ANTIKEN MUSEUM
VIAMUS – DAS VIRTUELLE ANTIKENMUSEUM
start

Access example using emulated client

Re-enact on demand

- complete environment citable (HDL)
- efficent preservation strategy for complex networked environments
-computing costs scale with access demand
- building blocks for a virtual server ecosystem

Instructions

1. Scan QR code
2. Wait until the system has started
3. Start browser and enter: **viamus.de:8003**

Sustainable, justifiable, repeatable: A digital preservation strategy using metrics-based (re)appraisal

Brent West
University of Illinois
506 S Wright St, M/C 359
Urbana, IL 61801 USA
+1 217-265-9190
bmwest@uillinois.edu

Joanne Kaczmarek
University of Illinois
506 S Wright St, M/C 359
Urbana, IL 61801 USA
+1 217-333-6834
jkaczmar@illinois.edu

Jordan Phoenix
University of Illinois
506 S Wright St, M/C 359
Urbana, IL 61801 USA
+1 217-300-1915
jphoenix@uillinois.edu

ABSTRACT

Appraisal has long been a source of intense debate in the archival community. Digital collections, affordable storage costs, and software tools now offer the opportunity to enhance appraisal strategies and move the archival appraisal discussion productively forward. In this poster, we propose an iterative, technology-assisted, metrics-based approach to appraisal as part of a digital preservation strategy. Developed in concert with a Capstone-based project to preserve email messages of enduring value at the University of Illinois, this multimodal approach integrates various traditional appraisal techniques with business performance metrics for the purpose of achieving growth that is sustainable using justifiable appraisal decisions, applying repeatable processes, and ultimately establishing measurable institutional value.

General Terms

Preservation strategies and workflows, theory of digital preservation, case studies and best practice.

Keywords

Archives, appraisal, business process improvement, capstone, digital preservation, metrics, mplp, reappraisal, sustainability.

1. BACKGROUND

The University of Illinois Records and Information Management Services (RIMS) office is currently engaged in a project to help its campus' archivists preserve email messages of enduring value beginning with those of its senior administrators. The underlying assumption is that the email messages of senior administrators are the modern equivalent of the traditional subject or general correspondence files which have long been determined to have enduring value for administrators and researchers alike. However, email presents unique challenges to archival accessioning including volume, file format, links, attachments, use for both personal and official communications, conversation threads, inconsistent filing, sensitive content, and ease of search and copying.

The volume of email content, mix of personal and professional usage, and an inability to rely upon diverse administrators to consistently identify messages of enduring value led the RIMS project to explore the Capstone approach developed by the United States National Archives and Records Administration's (NARA) [1]. The Capstone approach offers an option for agencies to capture and preserve most of the email from the accounts of officials at or near the head of an agency without detailed consideration of the content. However, it is clear that far reaching preservation approaches can quickly become unsustainable if re-appraisal is not part of the process. While digital storage is relatively cheap, it is not free. In addition, the costs for processing and digital preservation workflows will increase as accessioned content increases. This increasing demand on limited resources places additional risk on high-value content. It is in this context that our metrics-based appraisal approach is proposed.

Upfront, the RIMS project plans to provide administrators with tools that can assist in making informed options to identify messages that are of a personal nature or that warrant access restrictions such as those containing sensitive information. Most messages will not be transferred to a digital archival repository managed by the University Archives until some period of time after the administrator has left their position. Once deposited, the messages are expected to be subject to a restriction period during which archivists would have the opportunity to apply archival processing techniques to the materials. Once materials are made available to researchers through usual and customary archival controls, usage statistics will be gathered to be included in a future re-appraisal stage of the digital content held by the University Archives.

2. APPRAISAL STRATEGIES

2.1 Traditional Appraisal

Traditional appraisal strategies for archival collections typically rely on the professional subjective opinion of an archivist based upon characteristics of the record such as current and anticipated use and functional value [2]. NARA's appraisal policy, for example, is fourteen pages in length and includes subjective guidance questions such as "How significant are the records for research?" [3]. An appraisal strategy that relies primarily on subjective evaluation can result in over-retention, underutilized holdings, and inconsistent guidance given to record creators. This can lead to archival holdings which are never accessed yet continue to consume archival resources while they fail to bring any value to the institution.

2.2 Proposed Metrics-Based Appraisal

An objective, justifiable appraisal process would benefit records creators, archivists, and researchers alike. How then can an archives sustainably curate its digital collections and enhance institutional value in both the short and long-term? The proposal is to minimize the subjective factors by intentionally over-accessioning the digital records. Once accessioned, tools would be applied to assist in filtering and gathering use statistics. Regardless of the initial appraisal method applied, a metrics-based reappraisal approach can align existing resources and holdings with the institutional needs. Technology-assisted digital preservation should afford the opportunity to focus less on acquisitions and more on outreach and programming.

How is this accomplished? By combining archival principles such as "More Product, Less Process" (MPLP) with business-driven records management and performance standards, archivists can relax appraisal strategies at the point of acquisition and instead incorporate a recurring reappraisal stage to the management of their holdings. The reappraisal stage would include data from use statistics over time coupled with a "value score" assigned by the archivist at the time of reappraisal. These metrics will allow archivists to gauge value based on interest shown in particular collections or series, as well as their professional assessment of the significance of the materials.

After a predefined period of time, records series or collections which have low use statistics coupled with low value scores would be placed on a "watch list" for a period of a few years. While materials are on the watch list, if archivists feel the content was overlooked by users and warrants continued retention, they may engage in targeted outreach and programming to foster interest. At the end of the watch list period, records series that continued to be underutilized would be rated as having limited value and de-accessioned or relegated to an inactive status. In recognition of the archivists' professional judgment and the need to retain a human factor in any appraisal approach, at their discretion archivists could retain a subset of de-accessioned/ inactive status materials. The specific period of time between reappraisal decisions should be tailored to the specific needs and resources of the institution.

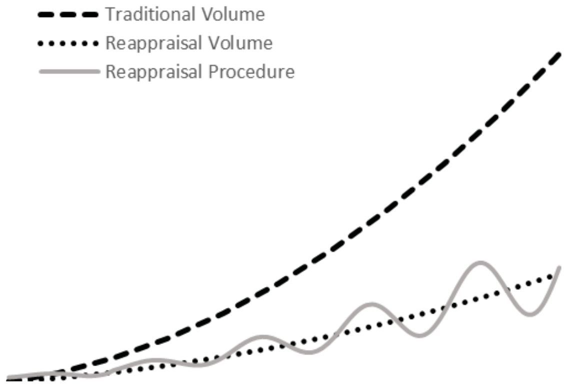


Figure 1. Traditional versus reappraisal volume growth over time.

2.3 Benefits

A metrics-based appraisal strategy provides many advantages. While not a perfect solution for every institution, it represents an approach which is workable while being more sustainable in an environment of explosive growth and limited resources as shown in Figure 1. In particular, mixed-value content such as email which is currently lost at an unprecedented scale can be more easily appraised by archivists and record creators alike. Administrative support of the archives may be more forthcoming if administrators find direct value in its contents. For instance, material not classically valued by archivists but which supports business functions and continuity may be easier to justify including in the digital repository. Metrics-based reappraisal allows collections to self-distill in an organic yet controlled manner that is reasonably consistent and repeatable between archivists. Metrics also provide support for both digital and analog preservation strategies and demonstrable value of the return on investment to the institution.

2.4 Sample Strategy

As an example, a preliminary appraisal decision may call for a ten year period during which use statistics are collected. If a particular series fell below the twentieth percentile of access over the ten year period, it would be placed on a watch list for the next ten years. During this time, the archivist can choose to conduct programming to promote interest in a topic related to the series. After twenty years, if the file still remains below the twentieth percentile, the archivist would prepare to de-accession the materials. At this point the archivist may elect to use his/her discretion to retain a percentage of the underutilized materials in the series due to its unique characteristics or some other clearly articulated criteria based on her/his professional judgment.

3. NEXT STEPS

We propose a metrics-based approach to appraisal as a case study for a project to preserve email messages of enduring value at the University of Illinois. We seek to refine the strategy through practical application and to provide lessons learned on its effectiveness for others that wish to implement a similar strategy within their organization. We will also explore simulated applications with other digital repository content to develop experience with a more broad range of digital materials.

4. REFERENCES

[1] U.S. National Archives and Records Administration. 2013. NARA Bulletin 2013-02. http://www.archivess.gov/records-mgmt/bulletins/2013/2013-02.html.

[2] Pearce-Moses, R. 2005. A *Glossary of Archival and Records Terminology*. Society of American Archivists.

[3] U.S. National Archives and Records Administration. 2007. Strategic Directions: Appraisal Policy. http://www.archivess.gov/records-mgmt/initiatives/appraisal.html.

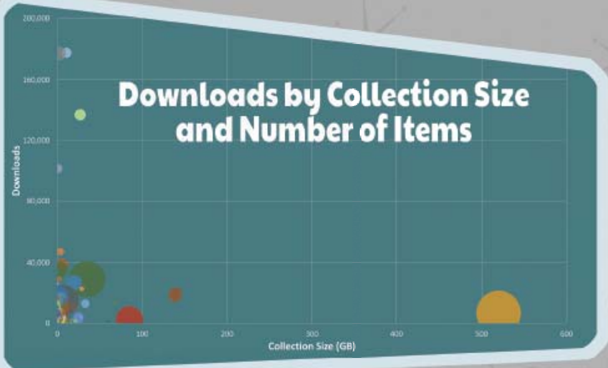
Sustainable, justifiable, repeatable:
Digital preservation using metrics based (re)appraisal

Problem Statement

Subjective appraisal can result in:

- Over-retention
- Underutilized holdings
- Inconsistent guidance
- Diversion of resources

Downloads by Collection Size and Number of Items

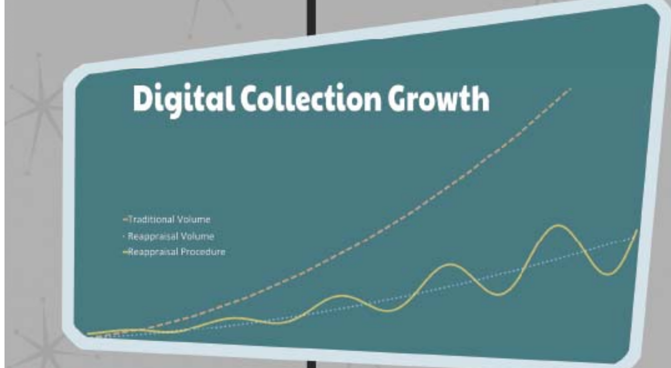


About Us

Records and Information Management Services (RIMS) helps the University of Illinois:

- Protect vital records
- Reduce legal liability
- Support the preservation of historic records
- Promote scholarship and teaching excellence
- Improve operational efficiencies

Digital Collection Growth



Metrics Based (re)Appraisal

MPLP meets RIM meets BPI

- Relaxed appraisal strategy
- Recurring reappraisal
- Usage statistics
- Archivist value score

Next Steps

Case study:

- Refine through practical application
  - University of Illinois email preservation project
- Provide lessons learned on its effectiveness
- Explore simulated applications with other digital repository content

http://go.uillinois.edu/rimsMBR

Work Flow

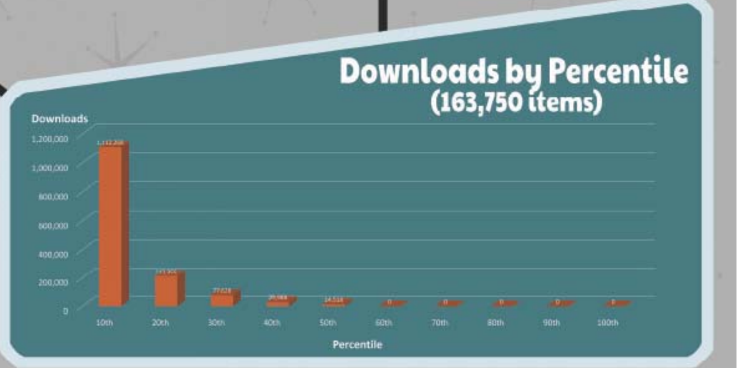
Example strategy:

- Collect usage statistics for 10 years
- Add <20th percentile series to watch list
- Promote series through programming
- Deaccession at 20 years if:
  - Still <20th percentile and
  - Value score 2 out of 5 or less

University of Illinois
Urbana-Champaign • Chicago • Springfield

Brent West - Joanne Kaczmarek - Jordan Phoenix

Downloads by Percentile
(163,750 items)



Presto4U- European Technology for Digital Audiovisual Media Preservation

Jacqui Gupta
BBC R&D
London
UK
+44 (0)3030 409638
jacqui.gupta@bbc.co.uk

ABSTRACT

This poster describes the actions, progress and research results of the Presto4U project, an EC funded two year coordination and support action which started in January 2013 focussing on the long term preservation of digital audiovisual media within a diverse range of Communities of Practice. The aim of the project is to identify useful results of research into digital audiovisual preservation and to raise awareness and improve the adoption of these both by technology and service providers as well as media owners. It will deliver new tools and services to connect different constituencies involved in AV Media preservation: expert users, who understand the problems and require technological solutions; researchers who can develop the fundamental knowledge; and technology providers who can commercialise research results as sustainable tools and services.

General Terms

communities, preservation strategies and workflows, digital preservation marketplace

Keywords

digital media preservation, communities of practice, standards, digital preservation marketplace, preservation tools, best practice

1. INTRODUCTION

The long-term preservation of digital audio-visual media presents a range of complex technological, organisational, economic and rights-related issues, which have been the subject of intensive research over the past fifteen years at national, European and international levels. Although good solutions are emerging, and there is a large body of expertise at a few specialist centres, it is very difficult for the great majority of media owners to gain access to advanced audio-visual preservation technologies. Presto4U aspires to raise awareness and improve the adoption of audio-visual preservation research results, both by service providers and media owners, and with a particular emphasis on meeting the needs of smaller collections, private sector media owners and new stakeholders.

2. AIMS AND OBJECTIVES

The project aims to create a series of Communities of Practice in the principal sub-sectors of audiovisual media preservation and develop a body of knowledge on the status of digital preservation practice, outstanding problems and needs for access to research results;

-identify useful results of research into digital audiovisual preservation;

-promote the take-up of promising research results by users, technology vendors and service providers, based on results of hands-on technology assessment, promotion of standards, analysis of economic and licensing models, and provision of brokering services;

-raise awareness of the need for audiovisual media preservation and disseminate information about project results; and

-evaluate the impact of the project and develop plans for long-term sustainability**.**

3. COMMUNITIES OF PRACTICE

Nine Communities of Practice have been identified and established in the field of AV preservation, each based on a shared concern, a shared set of problems and a common pursuit of technological solutions. By collaboratively sharing examples of the problems and challenges that their audiovisual collections face, they are raising awareness and contributing to knowledge creation and transfer of the issues and concerns specific to audiovisual preservation. Expert groups within each community of practice have been set up to provide a key exchange environment through meetings and networking events, pooling the available expertise between academic research, media, culture and industry sectors. Identifying the commonalities and differences across communities functions as a way of inspiring the research sector to develop products that better fit the needs of the different communities. The Communities are as follows each with a Community of Practice leader from a project partner:

- Music and Sound Archives (INA)
- TV, Radio and New Media broadcasting
- Video Production and Postproduction (TV2)
- Film Collections and Filmmakers (DFI)
- Video Art, Art Museums and Galleries (TATE)
- Footage Sales libraries (LUCE)
- Research and Scientific Collections (CNR)
- Learning & Teaching Repositories (KCL)
- Personal Collections (INA)

4. PROJECT RESEARCH OUTCOMES

4.1 Identification and Analysis of Research Outcomes

The core work will be a preservation research technology watch and assessment to identify potentially useful results of research into digital preservation and to track and map research projects, emerging commercial technologies and new technical approaches to establish their readiness for take-up, and to match the specific needs of each Community of Practice.

A research outcomes methodology will be developed to create a set of criteria, metrics and test datasets for analysis, comparison and assessment of these technological outputs and the results will be published through PrestoCentre. Outcomes of research initiatives that address issues related to rights will also be tracked and analysed including definition of formats for rights expression languages, models for representation of rights ontologies, new services for the management of rights information and technologies for enforcing appropriate use of rights.

4.2 Technology Transfer from Preservation Research

The project will promote the take-up of promising research results and encourage adoption of technologies emerging from digital preservation R&D that solve problems experienced by the Communities but which have not yet reached the market. It will investigate barriers to the uptake of research results, including licensing to vendors for productisation and ways of engaging new suppliers into the market place. Promoting technology standardisation and services is another task which will support the application of standards based tools and services by analysing audio-visual and preservation standards relevant to each Community of Practice, including upcoming specifications and the process for adoption. The outcome will be a set of standards recommendations and the creation of a Standards Register which will include a taxonomy of standards and guidelines on application of standards for technologies within each Community of Practice, providing evidence of adoption. Further outcomes will include a Tools Catalogue and a Market Place.

4.3 Impact and Sustainability

There will be an evaluation of the impact of the project on the uptake of research outcomes by the audio-visual media preservation sector, analysing feedback from the Communities of Practice and researchers and conducting impact and satisfaction surveys. It will also develop plans for long term sustainability by gathering and developing economic models and business plans for the long term maintenance of activities by PrestoCentre[1]. PrestoCentre will be open to developing the scope of the Communities of Practice as opportunities present themselves, for example, the work of the Pocos project on complex media objects[2].

5. ACKNOWLEDGEMENTS

This work was supported by the European Commission Seventh Framework Programme under Grant Agreement Number FP7-600845

Project Partners:- Beeld en Geluid-Netherlands (B&G),Danish Film Insitute-Denmark (DFI),Tate Gallery- UK (Tate), TV2-Denmark (TV2), Cinecitta Luce-Italy (Luce), British Broadcasting Corporation-UK(BBC),Institut National de l’Audiovisuel (INA) Radiotelevisione Italiana-Italy (RAI), Moving Media Ltd-Ireland, Kings College London-UK (KCL), Consiglio Nazionale Delle Richerche-Italy (CNR),University of Southampton IT Innovation-UK, Eurix-SRL Italy, Joanneum Research-Austria(JRS)

PRESTO4U
Presto4U is a two-year project supported by a consortium of fourteen partners from seven EU countries covering a wide range of preservation expertise based on extensive research, multiple Communities of Practice, and centres specialising in technology transfer between research and industry. The project receives funding from the European Commission's Seventh Framework Programme [3].

6. REFERENCES

[1] https//www.prestocentre.eu

[2] https//www.pocos.org

[3] https//www.presto4u.eu

Presto4U

European Technology for Digital Audiovisual Media Technology

Cultural and media organisations struggle to find technologies and tools for digital audiovisual preservation, and to evaluate their suitability. Presto4U identifies and evaluates such technologies and tools and promotes their adoption by archives and technology and service providers.

Our Mission

Presto4U aims to identify useful results of research into digital audiovisual preservation and to raise awareness and improve the adoption of these by both technology and service providers as well as media owners. The project will deliver new tools and services to connect the different constituencies involved in AV media preservation: expert users who understand the problems and require technological solutions; researchers who can develop the fundamental knowledge; and technology providers who can commercialise research results as sustainable tools and services.

www.prestocentre.org/4u

INTRODUCTION

The project aims to create a series of Communities of Practice in the principal sub-sectors of audiovisual media preservation and develop a body of knowledge on the status of digital preservation practice, outstanding problems and needs for access to research results;

-identify useful results of research into digital audiovisual preservation;

-promote the take-up of promising research results by users, technology vendors and service providers, based on results of hands-on technology assessment, promotion of standards, analysis of economic and licensing models, and provision of brokering services;

-raise awareness of the need for audiovisual media preservation and disseminate information about project results; and
-evaluate the impact of the project and develop plans for long-term sustainability.

COMMUNITIES OF PRACTICE

Nine Communities of Practice have been identified and established in the field of AV preservation, each based on a shared concern, a shared set of problems and a common pursuit of technological solutions. By collaboratively sharing examples of the problems and challenges that their audiovisual collections face, they are raising awareness and contributing to knowledge creation and transfer of the issues and concerns specific to audiovisual preservation. Expert groups within each community of practice have been set up to provide a key exchange environment through meetings and networking events, pooling the available expertise between academic research, media, culture and industry sectors. Identifying the commonalities and differences across communities functions as a way of inspiring the research sector to develop products that better fit the needs of the different communities. The Communities are as follows each with a Community of Practice leader from a project partner:

- Music and Sound Archives (INA)
- TV, Radio and New Media broadcasting
- Video Production and Postproduction (TV2)
- Film Collections and Filmmakers (DFI)
- Video Art, Art Museums and Galleries (TATE)
- Footage Sales libraries (LUCE)
- Research and Scientific Collections (CNR)
- Learning & Teaching Repositories (KCL)
- Personal Collections (INA)

Identification and Analysis of Research Outcomes

The core work will be a preservation research technology watch and assessment to identify potentially useful results of research into digital preservation and to track and map research projects, emerging commercial technologies and new technical approaches to establish their readiness for take-up, and to match the specific needs of each Community of Practice. A research outcomes methodology has been developed to create a set of criteria, metrics and test datasets for analysis, comparison and assessment of these technological outputs and the results will be published through PrestoCentre. Outcomes of research initiatives that address issues related to rights will also be tracked and analysed including definition of formats for rights expression languages, models for representation of rights ontologies, new services for the management of rights information and technologies for enforcing appropriate use of rights.

Technology Transfer from Preservation Research

The project will promote the take-up of promising research results and encourage adoption of technologies emerging from digital preservation R&D that solve problems experienced by the Communities but which have not yet reached the market. It will investigate barriers to the uptake of research results, including licensing to vendors for productisation and ways of engaging new suppliers into the market place. Promoting technology standardisation and services is another task which will support the application of standards based tools and services by analysing audio-visual and preservation standards relevant to each Community of Practice, including upcoming specifications and the process for adoption. The outcome will be a set of standards recommendations and the creation of a Standards Register which will include a taxonomy of standards and guidelines on application of standards for technologies within each Community of Practice, providing evidence of adoption. Further outcomes will include a Tools Catalogue and a Market Place.

Impact and Sustainability

There will be an evaluation of the impact of the project on the uptake of research outcomes by the audio-visual media preservation sector, analysing feedback from the Communities of Practice and researchers and conducting impact and satisfaction surveys. It will also develop plans for long term sustainability by gathering and developing economic models and business plans for the long term maintenance of activities by PrestoCentre[1]. PrestoCentre will be open to developing the scope of the Communities of Practice as opportunities present themselves, for example, the work of the Pocos project on complex media objects[2].

References

[1] https//www.prestocentre.eu
[2] https//www.pocos.org

Presto4U services

Whilst nurturing the Communities of Practice, the Presto4U project is also creating tools and resources to help the communities achieve their long-term digital preservation mission.

PrestoCentre Library
A wide range of free downloadable resources have been collated within the PrestoCentre Library to fulfil the various needs of users. Different resources for various levels of learning.
www.prestocentre.org/library

Webinars
Presto4U offers an on-going free webinar series on diverse topics related to audiovisual digitisation and digital preservation. Each webinar is focussed on a specific topic and hosted by experts within the field.
www.prestocentre.org/4u/publication-services

Standards Register
The Standards Register incorporates information on standards for content and metadata used across all communities involved in audiovisual digital preservation, with further input from experts in audiovisual preservation.
www.prestocentre.org/standards

Preservathons
PrestoCentre Preservathons are two-day hands-on events developed around main themes and challenges in audiovisual digitisation, preservation and long-term access.

Tools Catalogue
As well as open source and commercial tools, the catalogue incorporates tools and other emerging research results used at different stages of the lifecycle in long-term preservation of digital audiovisual media.
www.prestocentre.org/library/toolscatalogue

Market Place
This will be the place where users can express their audiovisual preservation needs and be presented with tailored solutions. Presto4U is currently scoping functional requirements and technical build of the brokerage environment for launch later this year.

The Presto4U project receives funding from the European Commission Seventh Framework Programme (FP7). Grant Agreement Number is FP7- 600845
The project will run from the 1 January 2013 - 31 December 2014.

Quality Assurance Tools for Digital Repositories

Roman Graf
Ross King
AIT - Austrian Institute of Technology GmbH
Donau-City-Strasse 1
Vienna, Austria
{roman.graf,ross.king}@ait.ac.at

ABSTRACT

Digitization workflows for automatic acquisition of image collections are susceptible to errors and require quality assurance. This paper presents a quality assurance tool suite for long term preservation. These tools support decision making for blank pages, cropping errors, mistakenly appearing fingers in scans and accurate duplicate detection in document image collections. The important contribution of this work is a definition of the quality assurance workflow and its automatic computation. The goal is to create a reliable tool suite that is based on image processing techniques.

Categories and Subject Descriptors

H.3.7 [**Digital Libraries**]: System issues

General Terms

preservation strategies and workflows

Keywords

digital preservation, quality assurance, image processing, information integration

1. INTRODUCTION

Within the last decade, significant effort has been invested in digitisation projects. Many large-scale digitization projects are running in digital libraries and archives and in public-private partnerships between cultural heritage institutions and industrial partners. The overall production in these projects has reached a level where a comprehensive manual audit of image quality of all digitized material would be neither feasible nor affordable. Nevertheless, cultural heritage institutions are facing the challenge of assuring adequate quality of document image collections that may comprise millions of books, newspapers and journals with hundreds of documents in each book. Quality assurance tools that aid the detection of possible quality issues are required. The material used in our experimental setup has been digitized in the context of Austrian Books Online, a public private partnership of the Austrian National Library with Google. In this partnership the Austrian National Library digitises and puts online its historical book holdings ranging from the 16th to 19th century with a scope of 600,000 books (see [1]). The project includes aspects ranging from digitisation preparation and logistics to quality assurance and online-access of the digitized items. Especially the quality assurance presents a challenge where automatic and semi-automatic tools are required to facilitate the quality assurance processes for the vast range and amount of material (described in [2]). The main contribution of this paper is the development of a DIGLIB QA Suite for the analysis of digital document collections and for reasoning about analysed data.

2. QUALITY ASSURANCE TOOLS

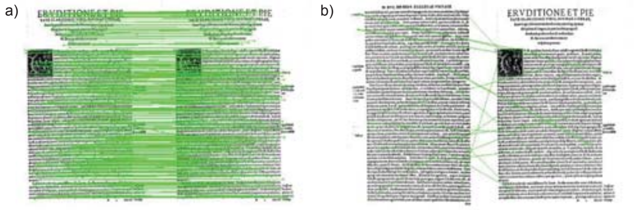


Figure 1: Samples of evaluation results from book identifier 151694702 (Austrian National Library) for duplicate detection with SIFT feature matching approach: (a) similar pages with 419 matches, (b) different pages with 19 matches.

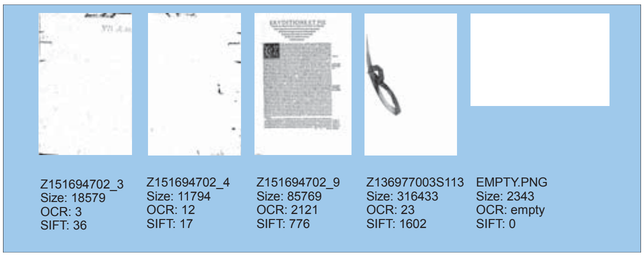


Figure 2: Selected samples of blank pages in digital collections from different sources with associated file name, file size, OCR and scale-invariant feature transform (SIFT) analysis result.

The suite includes four tools. The *matchbox* tool [3] for accurate duplicate detection in document image collections is a modern quality analysis tool based on Scale-Invariant Feature Transform (SIFT) feature extraction (see Figure 1). The *blank page detection* tool [4] that employs different image processing techniques and Optical Character Recognition (OCR) (see Figure 2). The *finger detection* tool [5] for automatic detection of fingers that mistakenly appear in scans from digitized image collections. This tool uses modern image processing techniques for edge detection, local image information extraction and its analysis for reasoning on scan quality (see Figure 3). The *cropping error detection* tool supports the analysis of digital collections (e.g. JPG, PNG files) for detecting common cropping problems such as text shifted to the edge of the image, unwanted page borders, or unwanted text from a previous page on the image (see Figure 4).

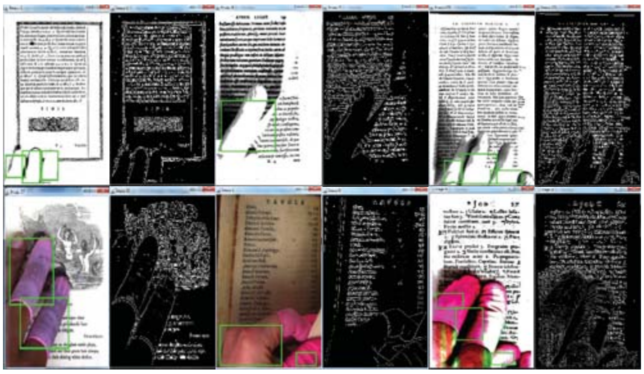


Figure 3: Positive detections of fingers on scans with associated edges where suspected areas are marked by green rectangles.

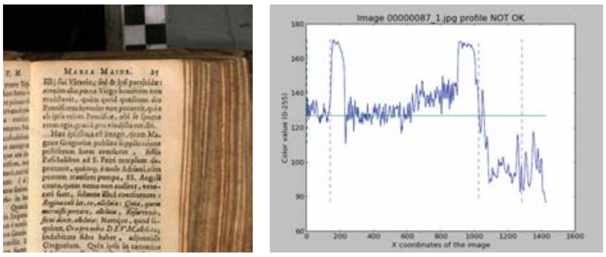


Figure 4: Cropping detection sample.

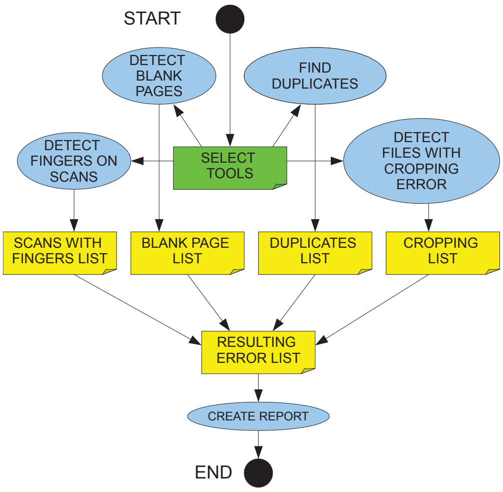


Figure 5: The workflow for the DIGLIB QA tool suite.

3. THE ERROR DETECTION PROCESS

The presented tools cover multiple error scenarios. The main use case for *matchbox* tool is a detection of the duplicated documents. Blank pages in a collection may address failure in a scanning process. Fingers should not be visible on the scans. The use cases for cropping errors are: text shifted to the edge of the image; unwanted page borders and unwanted text from the previous page on the image. Figure 5 presents the quality analysis workflow that employs different image processing tools for detection of errors and inaccuracies in digital document collections. This workflow includes the acquisition of local and global image descriptors, its analysis and an aggregation of resulting data in a single report for collection. The metadata and the selection criteria of digitization for preservation should be defined by an institutional expert for digital preservation. Selection criteria are dependent on particular collection types. Evaluation took place on an Intel Core i73520M 2.66GHz computer using Java 6.0 and Python 2.7 languages on Windows OS. The Relative Operating Characteristic (ROC) values for duplicate detection, cropping errors, blank pages and fingers on scans detection are represented by (0.013, 0.7), (0.001, 0.666), (0.007, 1.0) and (0.04, 0.844) points respectively. All these points are located very close to the so called perfect classification point (0, 1).

4. CONCLUSIONS

In this work we presented an approach for bringing together information automatically aggregated from different quality assurance tools regarding possible errors or inaccuracies in digital collection. The quality assurance tools for digital collections can help to ensure the quality of digitized collections and support managers of libraries and archives with regard to long-term digital preservation. As future work we plan to perform a statistical analysis of the automatically extracted information from the quality assurance tool and the qualitative analysis of the aggregated knowledge.

5. ACKNOWLEDGMENTS

This work was partially supported in part by the EU FP7 Project SCAPE (GA#270137) `www.scape-project.eu`.

6. REFERENCES

[1] M. Kaiser, “Putting 600,000 Books Online: The Large-Scale Digitisation Partnership between the Austrian National Library and Google,” *Liber Quarterly*, vol. 21, pp. 213–225, 2012.

[2] M. Kaiser and S. Majewski, “Austrian Books Online: Die Public Private Partnership Der Österreichischen Nationalbibliothek Mit Google,” *Bibliothek Forschung und Praxis*, vol. 37, pp. 197–208, 2013.

[3] R. Huber-Mörk and A. Schindler, “Quality assurance for document image collections in digital preservation,” in *Proc. of the 14th Intl. Conf. on ACIVS*, LNCS, (Brno, Czech Republic), Springer, September 4-7 2012.

[4] R. Graf, R. King, and S. Schlarb, “Blank page and duplicate detection for quality assurance of document image collections,” *APA CDAC 2014*, vol. 2014, pp. 87–97, February 5-6 2014.

[5] R. Graf and R. King, “Finger detection for quality assurance of digitized image collections,” *Archiving conference*, vol. 2013, pp. 122–125, 2013.

Tool Suite for Quality Assurance of Digital Repositories

Roman Graf and Ross King

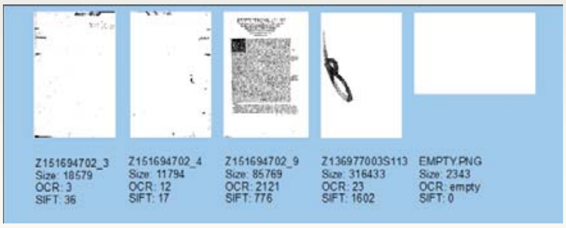
AIT Austrian Institute of Technology GmbH, Safety & Security Department, Vienna, Austria.

Roman.Graf@ait.ac.at and Ross.King@ait.ac.at

INTRODUCTION

Digitization workflows for automatic acquisition of image collections are susceptible to errors. Large document image collections in digital libraries require automatic quality assurance.

Fig 2. Selected samples of blank pages in digital collections from different sources with associated file name, file size, OCR and Scale-Invariant Feature Transform (SIFT) analysis result.



Z151694702_3	Z151694702_4	Z151694702_9	Z136977003S113	EMPTY.PNG
Size: 18579	Size: 11794	Size: 85769	Size: 316433	Size: 2343
OCR: 3	OCR: 12	OCR: 2121	OCR: 23	OCR: empty
SIFT: 36	SIFT: 17	SIFT: 776	SIFT: 1602	SIFT: 0

Fig 3. Positive detections of fingers on scans with associated edges where suspected areas are marked by green rectangles.

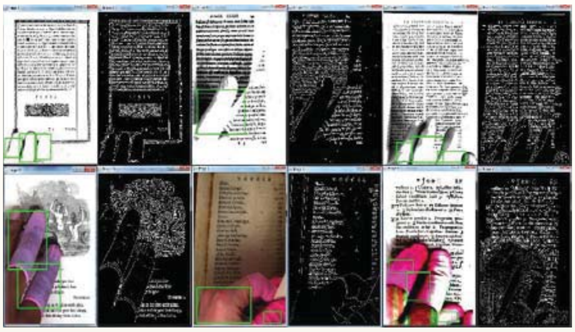
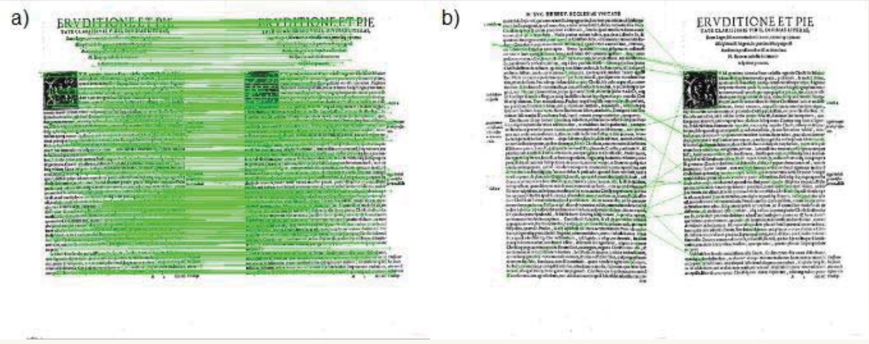


Fig 1. Evaluation results samples from book identifier 151694702 (Austrian National Library) for duplicate detection with SIFT feature matching approach: (a) similar pages with 419 matches, (b) different pages with 19 matches.



TOOLS

1.Matchbox tool [1] for accurate near duplicate detection in document image collections. A modern quality analysis tool based on SIFT feature extraction (see Figure 1).

2.Blank page detection tool [2] that employs different image processing techniques and OCR (see Figure 2).

3.Finger detection tool [3] for automatic detection of fingers that mistakenly appear in scans from digitized image collections. This tool uses modern image processing techniques for edge detection, local image information extraction and its analysis for reasoning on scan quality (see Figure 3).

4.Cropping error detection tool that supports the analysis of digital collections (e.g. JPG, PNG files) for detecting common cropping problems such as text shifted to the edge of the image, unwanted page borders, or unwanted text from a previous page on the image (see Figure 4).

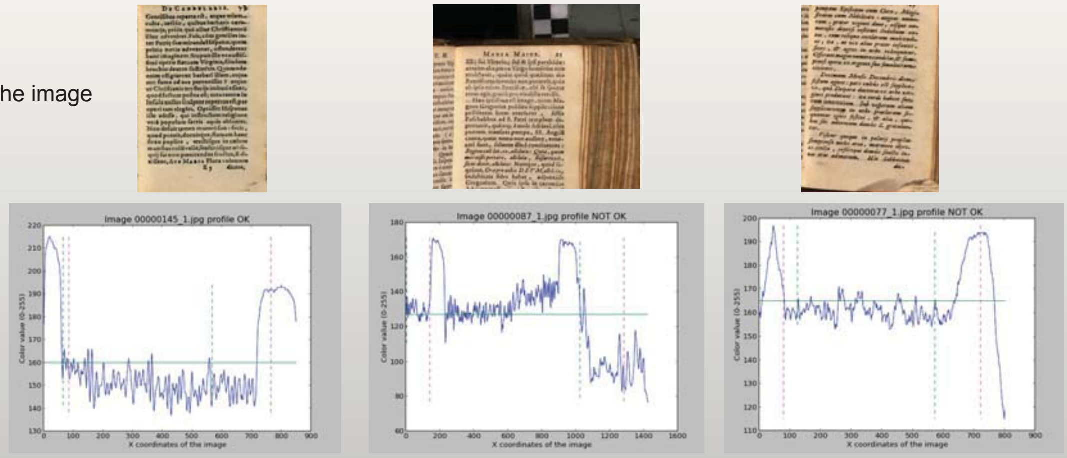
SOURCE CODE

1.Matchbox tool [http://openplanets.github.io/matchbox/]

2.Finger detection tool [http://openplanets.github.io/finger-detection-tool/]

3.Cropping error detection tool [http://openplanets.github.io/crop-detection-tool/]

Fig 4. Cropping detection samples



Use cases of cropping errors

- Text shifted to the edge of the image
- Unwanted page borders
- Unwanted text from previous page on the image

CONCLUSION

The quality assurance tools for digital collections can help to ensure the quality of digitized collections and support managers of libraries and archives with regard to long-term digital preservation.

REFERENCES

[1] Huber-Mörk, R., and Schindler, A. 2012. Quality assurance for document image collections in digital preservation. Proc. of the 14th Intl. Conf. on ACIVS (ACIVS 2012). LNCS, vol. 7517, pp. 108–119. Springer, Brno, Czech Republic.

[2] R. Graf, R. King, and S. Schlarb. Blank page and duplicate detection for quality assurance of document image collections. APA CDAC 2014, 2014:87-97, February 5-6 2014.

[3] Graf, R., and King, R. 2013. Finger Detection for Quality Assurance of Digitized Image Collections. Archiving conference. Volume 2013, April. 2013, 122-125.

This work was partially supported by the SCAPE Project. The SCAPE project is co-funded by the European Union under FP7 ICT-2009.4.1 (Grant Agreement number 270137).

Metadata Representation and Risk Management Framework for Preservation Processes in AV Archives

Werner Bailer
JOANNEUM RESEARCH – DIGITAL
Steyrergasse 17
8010 Graz, Austria
+43 316 876 1218
werner.bailer@joanneum.at

Martin Hall-May, Galina V. Veres
University of Southampton – IT Innovation Centre
Gamma House, Enterprise Road
SO16 7NS, Southampton, United Kingdom
+44 23 8059 8866
{mhm,gvv}@it-innovation.soton.ac.uk

ABSTRACT

This paper proposes an approach to assessing risks related to audiovisual (AV) preservation processes through gathering and representing metadata. We define a model for process metadata, which is interoperable with both business process models and other preservation metadata formats. A risk management framework is also suggested to help key decision makers to plan and execute preservation processes in a manner that reduces the risk of 'damage' to AV content. The framework uses a plan, do, check, act cycle to continuously improve the process based on risk measures and impact model. The process metadata serves as the interface between the steps in the framework and enables a unified approach to data gathering from the heterogeneous tools and devices used in an AV preservation workflow.

General Terms

Infrastructure, preservation strategies and workflows.

Keywords

Process metadata, business processes, risk management, risk assessment, simulation.

1. INTRODUCTION

Preservation processes for audiovisual content consist of complex workflows involving numerous interrelated activities performed by different tools and devices. Interoperable metadata throughout the entire workflow is a key prerequisite for performing, monitoring and analysing such preservation processes.

2. METADATA REPRESENTATION

For preservation purposes two types of metadata are most crucial: structural metadata (technical metadata needed to correctly interpret the stored essence) and preservation metadata (metadata for assessing the fixity, integrity, authenticity and quality of the object, as well as a documentation of the preservation actions applied). While the first is sufficiently covered by many existing formats, there is still a gap for representing preservation metadata for AV preservation processes. This paper focuses on the second type of metadata. Such processes as ingest, digitisation or migration can be quite complex, and heterogeneous workflows involve a number of different devices, software tools/ systems and users. We propose a metadata model for documenting the procedures applied to multimedia content in a preservation process together with tools, their parameters and operators involved. These metadata can be used for different applications, such as automatically adapting preservation and restoration workflows/tools, or collecting data for the assessment/simulation of risks related to these processes.

The scope of the preservation process metadata model is to document the history of creation and processing steps used, as well as their parameters. The model represents the preservation actions that were actually applied, i.e. a linear sequence of activities with the option to have a hierarchy for grouping activities. It supports a set of specific types of activities in the model (e.g., digitisation) with possible further specialisations (e.g. film scanning) in order to improve interoperability between preservation systems. The model also describes the parameters of these activities. There is a core set of well-defined properties together with their types, which store the value used when processing the item described. In addition, a key/value structure for supporting extensions is provided.

The model is designed around three main groups of entities: Content entities (DigitalItems, their Components and related Resources), Activities and Operators (Agent, Tool) and their properties. The *DigitalItem* represents an intellectual/editorial entity to be preserved. This entity has been borrowed from the MPEG-21 Digital Item Declaration (DID) model [1]. A DigitalItem aggregates other DigitalItems, such as the representations of an intellectual/editorial entity and the essences constituting the representation, and *Components*, such as the bitstreams of an essence. A *Component* is the binding of a resource to a set of metadata. It is not an item in itself, but a building block of items. It aggregates *Resources*, which are individually identifiable content files or streams in a container. A resource may also potentially be a physical object. All resources shall be locatable via an unambiguous address. Specialised subclasses of DigitalItem (such as supported in MPEG-21 and PREMIS [3]) can be optionally added, but are not needed for the purpose of describing preservation history. The model allows describing DigitalItems and Components without related Resources, which is useful for describing preservation activities that failed and left no trace in form of essence, but have to be documented for risk assessment.

An *Activity* is an action in the lifecycle of the content item which creates, uses or modifies a DigitalItem. Activities may be composed of other fine-grained Activities. Activities have start and end times, and their inputs/outputs are identified. This enables the reconstruction of the execution order and dependencies without an explicit description of serial or parallel activities and without having specific start/end events. Thus we achieve a simpler representation than in process models such as BPMN [2]. Having a generic activity and no discrimination into tasks and sub-processes harmonises handling preservation process descriptions with different granularity. Types of activities are modelled by reference to a controlled vocabulary, rather than defining the classes in the model.

An *Operator* is an entity contributing to the completion of an Activity by performing it or being used to perform it. The type of involvement is further specified by the Operator's role attribute. An Operator is either an *Agent* (a person or organisation involved in performing an activity) or a Tool (a device or software involved in performing an activity). The description of tools includes parameters and resource usage information. Operators may act on behalf of other Operators (e.g., Tools being used by Agents).

The metadata model constitutes a subset of the MPEG MP-AF data model described in [5].

3. RISK MANAGEMENT FRAMEWORK

We propose a risk management framework to help key decision makers to plan and execute preservation processes in a manner that reduces the risk of 'damage' to AV content. Damage is considered to be any degradation of the value of the AV content with respect to its intended use by a designated community that arises from the process of ingesting, storing, migrating, transferring or accessing the content.

This risk management framework relies on a repetitive procedure of planning and simulation of preservation processes, adapting and executing them, and gathering data from the execution for updating the risk model and simulation. Data gathering requires breaking down the process model, which contains all possible execution paths, into the sequence of actions that have actually been executed. Then the data are collected from configuration and execution logs of the individual tools. These data have to include not only operational information, but also risks-related knowledge. This knowledge consists of identified risks and their frequency of occurrence, their negative consequences and effects on assets (AV content), any controls dealing with the risks and their associated time and cost.

A cycle of continuous process improvement is proposed, which involves the following steps: plan, do, check, act. The basis of planning decisions is a simulated business process, representing the critical activities, tools and properties of the key preservation workflows (ingest, migration and access). The critical part to such a risk management approach is to ensure that the models and simulations of business processes used for planning decisions are kept consistent with the actual execution.

Most tools available for business process modelling are generic, offering no particular guide to the modeller. We use a controlled vocabulary to help to design the workflow, describe risks and thereby synchronise with the execution model. It also allows us to relate data gathered from the executing process to the activity in the workflow and to determine when and how risk measures are being breached. Three risk measures are suggested for preservation processes: Expected loss (mean of negative consequences (NC) which can occur in a given process), Value at Risk (minimum NC incurred in α% of the worst cases in a given process) and Conditional Value at Risk (mean of NC incurred in α% of the worst cases). These risks measures can be calculated allowing both propagation of risks through the preservation process and usage of controls to deal with risks occurred.

The metadata model is the interface between simulation and execution, as it allows us to map from abstract preservation activities, tools and their significant properties to and from their actual implementation. Metadata on process execution can be gathered for statistical analysis, and allows us to monitor preservation workflows in a manner that is consistent with planning models.

The purpose of the risk management framework is to allow the archive decision-makers to balance the cost and time involved in avoiding and mitigating risks with the risk reduction achieved by deploying 'controls' in the business process. By closing the loop between simulation and execution, the reliability and accuracy of the data used to drive planning decisions is improved, which is critical to justify any additional expenditure for uncertain future gains (i.e. long-term access to content).

To classify the impact of risks in digital preservation, we use the Simple Property-Oriented Threat Model (SPOT) as an impact model for Risk Assessment. The SPOT model [4] defines six essential properties of digital preservation: Availability, Identity, Persistence, Renderability, Understandability, and Authenticity.

The implemented demonstrator uses the metadata model to represent the data gathered from process definitions and execution logs and runs simulations using the risk assessment.

4. CONCLUSION

The proposed approach enables decision makers in AV preservation to make their decisions based on information about the risks involved. The risks can be assessed and simulated not only on estimates but on actual data gathered from the execution of preservation processes. This will provide a much more realistic and reliable assessment of risks and thus allow the risks of a damage to audiovisual content to be better managed.

5. ACKNOWLEDGMENTS

This work has been funded partially under the 7th Framework Programme of the European Union within the ICT project "DAVID" (ICT FP7 600827).

6. REFERENCES

[1] ISO/IEC 21000-2, Information technology – Multimedia framework (MPEG-21) – Part 2: Digital Item Declaration

[2] Object Management Group Business Process Model and Notation. http://www.bpmn.org/

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

[4] Vermaaten, S., Lavoie, B., and Caplan, P. 2012. Identifying Threats to Successful Digital Preservation: the SPOT Model for Risk Assessment, *D-Lib Magazine*. 18, 9/10, 2012.

[5] Allasia, W., Bailer, W., Gordea, S. and Chang, W. A Novel Metadata Standard for Multimedia Preservation, *Proceedings of iPres*, Oct. 2014.

Metadata Representation and Risk Management Framework for Preservation Processes in Audiovisual Archives

Werner Bailer[1], Martin Hall-May[2], Galina V. Veres[2]

[1]JOANNEUM RESEARCH
Forschungsgesellschaft mbH

DIGITAL
Institute for Information and
Communication Technologies

Steyrergasse 17
8010 Graz, Austria

Tel. +43 316 876-5000
Fax +43 316 876-5010

digital@joanneum.at
www.joanneum.at/digital

[2] University of Southampton
IT Innovation Centre

Gamma House, Enterprise Road
Southampton SO16 7NS
United Kingdom

Introduction

- Preservation processes for audiovisual content consist of complex workflows
- Activities are performed by different tools and devices
- Planning and improving workflows requires assessment of related risks
- Interoperable metadata is a key prerequisite for performing, monitoring and analysing such workflows

Metadata Representation

Two types of metadata are crucial

- Structural metadata: technical metadata
- Preservation metadata: assessing fixity, integrity, authenticity and quality, documentation of the preservation actions

DAVID metadata model

- Focus on documenting preservation activities applied to (digital) items
- Tools and agents involved, their attributes
- Represent data needed for risk assessment
- Compatibility with business process models (e.g. BPMN)
- Contributed to MPEG Multimedia Preservation Application Format (MP-AF)

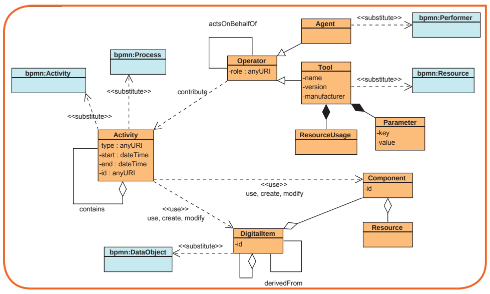


Figure 1: Entities of the preservation data model.

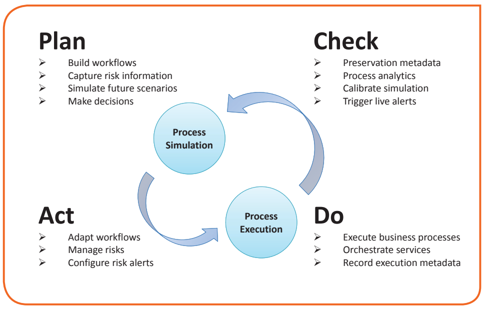
Risk Management Framework

Proposed cycle of continuous process improvement: plan, do, check, act

- Define objectives of risk management for digital preservation in archives
- Identify controls dealing with risks and any associated costs and time
- Analyse and classify risks according to an impact model (SPOT model)

Risk measures

- Expected loss (E): average magnitude (mean) of negative consequences
- Value at Risk (VaR): minimum negative consequence incurred in α% of worst cases
- Conditional Value at Risk (CVaR): expected negative consequence incurred in α% of worst cases



Plan

- Build workflows
- Capture risk information
- Simulate future scenarios
- Make decisions

Check

- Preservation metadata
- Process analytics
- Calibrate simulation
- Trigger live alerts

Process Simulation

Process Execution

Act

- Adapt workflows
- Manage risks
- Configure risk alerts

Do

- Execute business processes
- Orchestrate services
- Record execution metadata

Figure 2: Continuous business process improvement through monitoring and simulation.

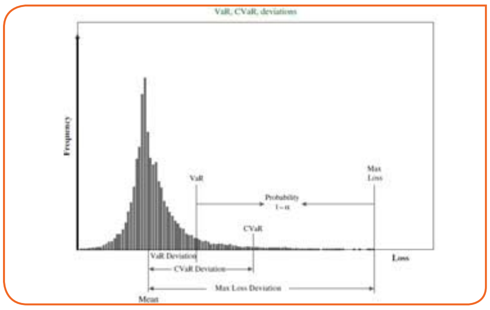


Figure 3: VaR, VaR deviation, CVaR, CVaR deviation, Maximum Loss and Maximum Loss Deviation (from [Sarykalin et al., 2008]).

Data Gathering

- Use the proposed metadata model as an interoperable representation of information from different tools
- Gather data from configuration, workflow engines and logs
- Include data about choices in workflow, exception handling and planned but not executed activities

Acknowledgements:

This work has been funded partially under the 7th Framework Programme of the European Union within the ICT project “DAVID” (http://www.david-preservation.eu, ICT FP7 600827).

ROHub – A Digital Library for Sharing and Preserving Research Objects

Raul Palma, Cezary Mazurek
Piotr Hołubowicz[i]
Poznan Supercomputing and Networking Center
Poznan, Poland
(+48) 618582161
[rpalma,mazurek]
@man.poznan.pl

Oscar Corcho
Ontology Engineering Group,
Universidad Politécnica de Madrid
Madrid, Spain
(+34) 913366605
ocorcho@fi.upm.es

José Manuel Gómez-Pérez
iSOCO
Madrid, Spain

(+34) 913349797
jmgomez@isoco.com

ABSTRACT

ROHub is a digital library system, enhanced with Semantic Web technologies, which supports the storage, lifecycle management, sharing and preservation of research objects - semantic aggregations of related scientific resources, their annotations and research context. ROHub includes a set of features to help scientists throughout the research lifecycle to create and maintain high-quality research objects that can be interpreted and reproduced in the future, including quality assessment, evolution management, navigation through provenance information and monitoring features. It provides a set of RESTful APIs along with a Web Interface for users and developers. A demo installation is available at: www.rohub.org.

General Terms

Infrastructure, specialist content types.

Keywords

Methods, preservation, semantic, aggregation, research objects

1. INTRODUCTION

Digital Library systems collect, manage and preserve digital content, with a measurable quality and according to codified policies [1]. These systems have been traditionally focused on the preservation of data and content of rather static nature, i.e., documents, images, datasets. However, research in data-intensive science, conducted in increasingly digital environments, has led to the emergence of new types of content and artefacts [2], such as computational methods that also have a dynamic dimension (i.e. they are executable). For instance, scientific workflows are executable descriptions of scientific procedures that define sequences of computational steps in automated data analysis.

Hence, in order to share and preserve research findings, we need to consider not only the data used and produced, but also the methods employed, and the research context in which these artefacts were conceived. Moreover, in order to enable the reusability and reproducibility of the associated investigations, we need to provide access to all these related artefacts, their research context, as well as information about the usage and provenance of these resources. Similarly, in order to capture the dynamic aspects of these resources, we need information about their evolution and, in the case of computational methods, about their executions.

Research objects (ROs) provide a container for all these associated artefacts. They are aggregating objects that bundle together experimental resources that are essential to a computational scientific study or investigation, along with semantic annotations on the bundle or the resources needed for the understanding and interpretation of the scientific outcomes. The RO model [3] provides the means for capturing and describing such objects, their provenance and lifecycle, facilitating the reusability and reproducibility of the associated experiments. The model consists of the core RO ontology[1], which provides the basic structure for the description of aggregated resources and annotations on those resources, and extensions for describing evolution aspects and experiments involving scientific workflows. Hence, ROs can help scientists in sharing research findings, but scientists also need the appropriate technological support enabling them to create, manage, publish and preserve these objects.

2. ROHub

ROHub is a digital library system supporting the storage, lifecycle management, sharing and preservation of research findings via ROs. It includes different features to help scientists throughout the research lifecycle: (i) to create and maintain ROs compliant with predefined quality requirements so that they can be interpreted and reproduced in the future; (ii) to collaborate along this process; (iii) to publish and search these objects and their associated metadata; (iv) to manage their evolution; and (v) to monitor and preserve them supporting their accessibility and reusability.

2.1 Interfaces

ROHub provides a set of REST APIs[2], the two primary ones being the RO API and the RO Evolution API. The RO API defines the formats and links used to create and maintain ROs in the digital library. It is aligned with the RO model, hence recognizing concepts such as aggregations, annotations and folders. The RO ontology is used to specify relations between different resources. ROHub supports content negotiation for metadata, including formats like RDF/XML, Turtle and TriG. The RO Evolution API defines the formats and links used to change the lifecycle stage of a RO, to create an immutable snapshot or archive from a mutable Live RO, as well as to retrieve their evolution provenance. The API follows the RO evolution model [3]. ROHub also provides a SPARQL endpoint, a Notification API, a Solr REST API, and a

[1] See http://wf4ever.github.io/ro/ and http://researchobject.org/

[2] APIs documentations available at: http://www.wf4ever-project.org/wiki/display/docs/Wf4Ever+service+APIs

User Management API, in addition to a Web interface, which exposes all functionalities to the users. The latter is the main interface for scientists and researchers to interact with ROHub.

2.2 Implementation

ROHub realizes the backbone services and interfaces of a software architecture for the preservation of ROs [4]. Internally, it has a modular structure that comprises access components, long-term preservation components and the controller that manages the flow of data. ROs are stored in the access repository once created, and periodically the new and/or modified ROs are pushed to the long-term preservation repository.

The access components are the storage backend and the semantic metadata triplestore. The storage backend can be based on dLibra[3], which provides file storage and retrieval functionalities, including file versioning and consistency checking, or it can use a built-in module for storing ROs directly in the filesystem.

The semantic metadata are additionally parsed and stored in a triplestore backed by Jena TDB[4]. The use of a triplestore offers a standard query mechanism for clients and provides a flexible mechanism for storing metadata about any component of a RO that is identifiable via a URI.

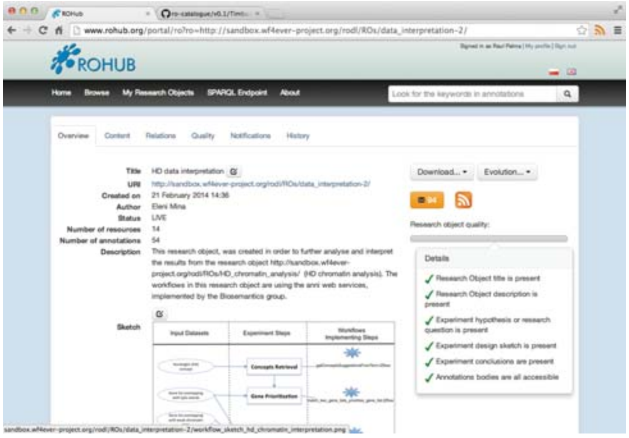
The long-term preservation component is built on dArceo[5], which stores ROs, including resources and annotations. Additionally, ROHub provides fixity checking and monitors the RO quality through time against a predefined set of requirements. If a change is detected, notifications are generated as Atom feeds.

2.3 Main functionalities

Create, manage and share ROs There are different methods for creating ROs in ROHub: (i) from scratch, adding resources progressively; (ii) by importing a pack of resources from other systems (currently myExperiment); (iii) from a ZIP file aggregating files and folders; (iv) by uploading local ROs from the command line using RO Manager Tool[6]. Resources can be added and annotated from the content panel that also shows the folder structure. ROHub provides different access modes to share the ROs: open, public or private. In the open mode, anyone with an account can visualise and edit the RO. In the public mode, everyone can visualise the RO, but only users with correct permissions can edit it. In private mode, only users with correct permissions can visualize and/or edit the RO. ROHub provides a keyword search box and a faceted search interface to find ROs, and a SPARQL endpoint to query RO metadata.

Assessing RO quality Users can visualise a progress bar on the RO overview panel (see Fig. 1), which shows the quality evaluation based on set of predefined basic RO requirements. When clicked, users can visualise further information about the RO compliance. Users can also get more information about the quality of the RO from the Quality panel, where they can choose from different templates to use as the basis for evaluating the RO.

Managing RO evolution From the RO overview panel, users can also create a snapshot (or release) of the current state of their RO, at any point in time, for sharing the current outcomes with colleagues, get feedback, send it to review, or to cite them.



www.rohub.org/portal/ro?ro=http://sandbox.wf4ever-project.org/rodl/ROs/data_interpretation-2/

Title HD data interpretation
URI http://sandbox.wf4ever-project.org/rodl/ROs/data_interpretation-2/
Created on 21 February 2014 14:36
Author Eleni Mina
Status LIVE
Number of resources 14
Number of annotations 54
Description This research object, was created in order to further analyse and interpret the results from the research object http://sandbox.wf4ever-project.org/rodl/ROs/HD_chromatin_analysis/ (HD chromatin analysis). The workflows in this research object are using the anni web services, implemented by the Biosemantics group.

Research object quality:

Details
✓ Research Object title is present
✓ Research Object description is present
✓ Experiment hypothesis or research question is present
✓ Experiment design sketch is present
✓ Experiment conclusions are present
✓ Annotations bodies are all accessible

sandbox.wf4ever-project.org/rodl/ROs/data_interpretation-2/workflow_sketch_hd_chromatin_interpretation.png

Figure 1 ROHub - RO overview panel

Similarly, when the research has concluded, they can release and preserve the outcomes for future references. ROHub keeps the versioning history of these snapshots, and calculates the changes from the previous one. Users can visualise the evolution of the RO from the History panel, and navigate through the RO snapshots.

Navigation of execution runs Scientists can aggregate any type of resources, including links to external resources and RO bundles, which are structured ZIP files representing self-contained ROs that facilitate their transfer and integration with 3rd party tools. Taverna, for example, can export provenance of workflow runs as RO Bundles. In ROHub, bundles are unpacked into nested ROs, exposing their full content and annotations. Hence, scientists can navigate through the inputs, outputs and intermediate values of the run, something potentially useful for future reproducibility.

Monitoring ROs ROHub includes monitoring features, such as fixity checking and RO quality, which generate notifications when changes are detected. This can help to detect and prevent, for instance, workflow decay, occurring when an external resource or service used by a workflow becomes unavailable or is behaving differently. Users can visualise changes in the RO, regarding the content and quality monitoring in the notification panel and they can subscribe to the atom feed to get automatic notifications.

3. ACKNOWLEDGMENTS

This work was supported by Wf4Ever EU project (FP7-270129). We thank Filip Wisniewski, Graham Klyne, Stian Soiland-Reyes and Wf4Ever users for their contributions and feedback.

4. REFERENCES

[1] Candela, L. et al. The DELOS Digital Library Reference Model Foundations for Digital Libraries. DELOS, Italy, Dec 2007

[2] De Roure, D. et al. Towards the preservation of scientific workflows. In Proceedings of the 8th International Conference on Preservation of Digital Objects (iPRES 2011). Nov 2011

[3] Belhajjame, K. et al. Workflow-centric research objects: First class citizens in scholarly discourse. In ESWC2012 Workshop on Semantic Publication (SePublica2012) May 2012

[4] Page, K. et al. From workflows to Research Objects: an architecture for preserving the semantics of science. In ISWC Workshop on Linked Science. Nov 2012

[3] http://dlab.psnc.pl/dlibra/

[4] http://jena.apache.org/

[5] http://dlab.psnc.pl/darceo/

[6] https://github.com/wf4ever/ro-manager

[i] Present address: Google, CA, USA. piotrhol@google.com

ROHub— A Digital Library for Sharing and Preserving Research Objects

Raúl Palma[1], *Piotr Hołubowicz*[1], *Oscar Corcho*[2], *José Manuel Gómez-Pérez*[3], *Cezary Mazurek*[1]
and the Wf4Ever consortium

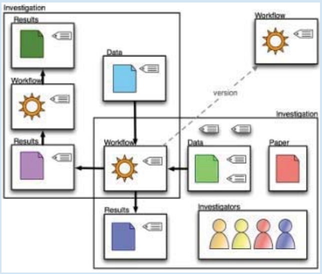
[1] Poznan Supercomputing and Networking Center, Poznan, Poland
[2] Ontology Engineering Group, Universidad Politécnica de Madrid, Madrid, Spain
[3] iSOCO, Madrid, Spain

URL	http://www.rohub.org/	http://www.rohub.org/rodl/
Demo video	http://youtu.be/TxW2wvreyoQ	http://youtu.be/gSEUswMmr8E
Source code	https://github.com/wf4ever/rodl	https://github.com/wf4ever/portal

ABSTRACT

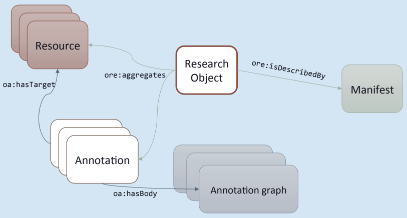
Research Objects (ROs) are semantic **aggregations** of related scientific resources, their **annotations** and research context. They provide the means to refer a **bundle** of research artifacts supporting an **investigation**, and the mechanisms to associate human and **machine-readable metadata** to these artefacts.

ROHub is a digital library system for ROs that supports their **storage**, **lifecycle management** and **preservation**. ROHub enables the **sharing of scientific findings** via ROs and includes features that help scientists throughout the research lifecycle to create and maintain **high-quality ROs** that can be **interpreted** and **reproduced** in the future.



RESEARCH OBJECT MODEL

A research object (RO) is described in an RDF **manifest** which lists the **aggregated resources** and their **annotations** as separate RDF graphs containing user annotations (*title, description, example value*), typing information (*hypothesis, workflow, input data, etc*) and automatically extracted metadata (*provenance, workflow structure*).



http://purl.org/wf4ever/model

The RO model consists of a core ontology and several extensions

- **ro core** provides the basic structure for the description of aggregated resources and annotations on those resources
- **roevo** captures the RO evolution, including the different stages during their lifecycle, the corresponding version along with their associated changes.
- **wfdesc** provides the vocabulary for the description of workflows
- **wfprov** provides the vocabulary for the description of workflow execution provenance

The ontologies for the RO Model are based on standards for aggregations (**OAI-ORE**), annotations (Annotation Ontology, W3C Open Annotation Core **OAC**) and provenance (W3C **PROV** ontology).

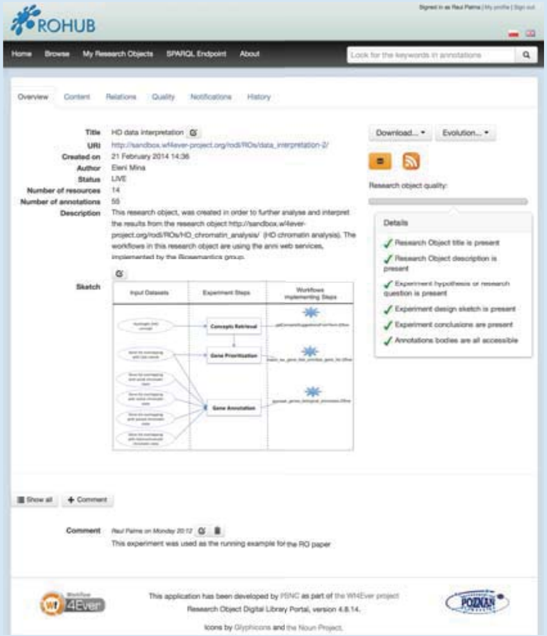
ROHUB INTERFACES

ROHub implements a set of open REST APIs, being the two primary ones:

- RO API - defines the formats and links used to create and maintain ROs in the digital library, according to RO model
- RO Evolution API - defines the formats and links used to change the lifecycle stage of a RO, as well as to retrieve their evolution provenance

ROHub also exposes a Notification API, User Management API, Access Control API, Solr API, and provides a SPARQL endpoint and a Web Interface

http://www.wf4ever-project.org/wiki/display/docs/Wf4Ever+service+APIs



Title: HD data interpretation
URI: http://sandbox.wf4ever-project.org/rodl/ROs/data_interpretation-2/
Created on: 21 February 2014 14:36
Author: Eleni Mina
Status: LIVE
Number of resources: 14
Number of annotations: 55
Description: This research object, was created in order to further analyse and interpret the results from the research object http://sandbox.wf4ever-project.org/rodl/ROs/HD_chromatin_analysis/ (HD chromatin analysis). The workflows in this research object are using the anni web services, implemented by the Biosemantics group.

Research object quality:

Details
- ✓ Research Object title is present
- ✓ Research Object description is present
- ✓ Experiment hypothesis or research question is present
- ✓ Experiment design sketch is present
- ✓ Experiment conclusions are present
- ✓ Annotations bodies are all accessible

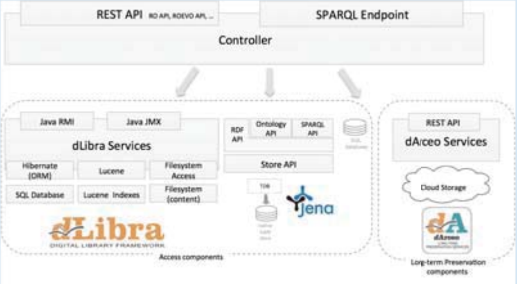
Comment: Raul Palma on Monday 2012
This experiment was used as the running example for the RO paper

This application has been developed by PSNC as part of the Wf4Ever project
Research Object Digital Library Portal, version 4.8.14.
Icons by Glyphicons and the Noun Project.

IMPLEMENTATION

ROHub has a **modular structure** that comprises **access components**, **long-term preservation components** and the **controller** that manages the flow of data. ROs are stored in the access repository once created, and periodically the new and/or modified ROs are pushed to the long-term **preservation repository**.

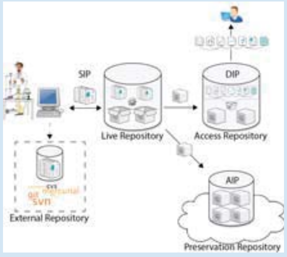
The storage backend can be based on **dLibra** (as shown below), or it can use a built-in module for storing ROs directly in the **filesystem**.



Preservation according to OAIS model

RO as Information Packages in OAIS

The archival system establishes three different storage repositories (or logical repositories) to support the preservation and access to preserved ROs



KEY FUNCTIONALITIES

a. **Create, manage and share ROs**: ROHub provides different methods for creating ROs and different **access modes** to share them: open, public or private. ROHub provides a **faceted search** interface, in addition to a keyword search box.

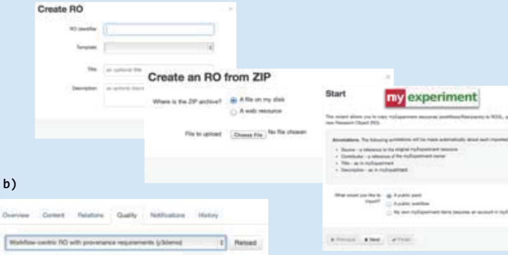
b. **Assessing RO quality**: In the Overview panel, ROHub shows a **progress bar** of the RO quality based on set of predefined basic RO requirements. Additional **quality information** can be visualized in the Quality panel.

c. **Managing RO evolution**: Users can create **RO snapshots** at any point in time, and release and **preserve** the RO when the research has concluded. Users can visualize the **evolution of the RO** from the History panel.

d. **Navigation of workflow run**: Scientists can **aggregate** any type of resource, including **links to external resources** and **RO bundles**, which are structured ZIP files representing self-contained ROs that facilitate their transfer and integration with 3rd party tools (e.g., Taverna)

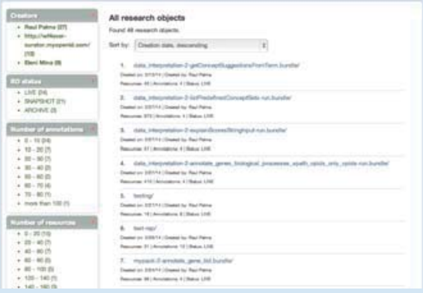
e. **Monitoring ROs**: ROHub includes monitoring features, such as fixity checking and **RO quality**, which generate **notifications** when changes are detected (e.g., workflow decay). Users can visualize this information in the Notification panel and they can subscribe to the **atom feed**.

a)

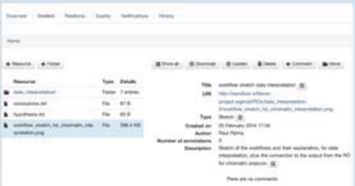


Where is the ZIP archive? ● A file on my disk ○ A web resource
File to upload: Choose File No file chosen

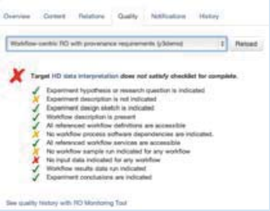
a)



a)



b)



Workflow-centric RO with provenance requirements (y3demo)

Target HD data interpretation does not satisfy checklist for complete.
- ✓ Experiment hypothesis or research question is indicated
- ✗ Experiment description is not indicated
- ✓ Experiment design sketch is indicated
- ✓ Workflow description is present
- ✓ All referenced workflow definitions are accessible
- ✗ No workflow process software dependencies are indicated.
- ✓ All referenced workflow services are accessible
- ✗ No workflow sample run indicated for any workflow
- ✗ No input data indicated for any workflow
- ✓ Workflow results data run indicated
- ✓ Experiment conclusions are indicated

See quality history with RO Monitoring Tool

c)

d)

e)

This work was enabled by the Wf4Ever project funded by the European Commission's 7th FWP (FP7-ICT-2007-6 270192), and partially by the Polish National Centre for Research and Development (NCBiR) under Grant No. SP/I/1/77065/10

A Biological Perspective on Digital Preservation

Michael J. Pocklington
Department of Genetics
University of Leicester
Leicester LE1 7RH, UK
m.pock@me.com

Anna Grit Eggers
Göttingen State and University Library
Georg August Universität
37070 Goettingen, Germany
eggers@sub.uni-goettingen.de

Fabio Corubolo
IPHS, University of Liverpool
Waterhouse Building
Brownlow Street
Liverpool L69 3GL, UK

Jens Ludwig
Göttingen State and University Library
Georg August Universität
37070 Goettingen, Germany
ludwig@sub.uni-goettingen.de

Mark Hedges
King's College London
Strand
London WC2R 2LS, UK
mark.hedges@kcl.ac.uk

Sándor Darányi
Swedish School of Library and Information Science
University of Borås
Allégatan 1, 50190 Borås, Sweden
sandor.daranyi@hb.se

ABSTRACT

Successful preservation of Digital Objects (DOs) ultimately demands a solid theoretical framework. Such a framework with a high degree of generality emerges by treating DOs as containers of functional genetic information, exactly as in the genomes of organisms. We observe that functionality links survival in organisms and utility in DOs. In both cases, functional information is identifiable in principle by the consequence of its ablation. In molecular biology, genetic ablations (mutations) and environmental ablations (experimental manipulations) are used to construct interaction maps fully representing organismic activity. The equivalent of such interaction maps are dependency networks for the use of DOs within their Digital Environment (DE). In the poster we will present early work on the application of the theoretical background. It includes first results from a case-study examining a software-based art preservation scenario (SBA) developed as part of the PERICLES FP7 project [1].

General Terms

Theory of digital preservation, preservation strategies and workflows.

Keywords

Digital ecosystems, digital preservation, niche, interaction map, significant environment information, sheer curation.

1. INTRODUCTION

Many active research programs exploit equivalences between biological objects and digital objects, up to and including, in the position taken by strong artificial life, the assumption of indistinguishability. The latter follows from the recognition that life is not dependent on any particular underlying medium, but is instead a property of evolving information-processing structures [2]. DP can not avoid such a viewpoint by internalisation and a retreat to technical issues, since it is embedded within policies and technologies that are themselves subject to the most rapid type of evolutionary change.

Scholars of culture have long debated the existence of autonomous informational processes in human society, and it seems likely that these become entangled with DOs, which inevitably evolve as technology advances. This brings issues for DP that may be best considered from a biological perspective. This is not merely a conceptual position: informational viruses and instant stock-trading algorithms can not be ignored, and seem to possess an autonomous evolutionary status. Despite repeated attempts at a generalised biological or Darwinian perspective of human organisational entities such as DOs, no consensus has been reached, even as to the best way to proceed.

DP is uniquely in a position of having to deal with DOs across the entire realm of human activity; they replicate, behave, consume resources, mutate, get selected, and evolve, demanding a meta-view of biology-based informational concepts. A key element for such a meta-view can be provided by systems biology. Systems biologists have found a way of visualising functions such as biochemical pathways and behaviours by interfering with genetic and environmental information, revealing the underlying structure of that information, in the form of genetic interaction maps. Similar methodology could be applied to DOs, to the benefit of their long term use, and reuse.

2. THEORETICAL FRAMEWORK

Underlying the existence of all biology is the specific context enabling organisms to survive, which is their niche. To call this an "environment" would be glib, as the niche is more than a regional container, but a specified provision of resources contingent upon the appropriate behaviour. We can operationally define information allowing survival by removing it one piece at a time. Traditionally we would call the removal of information "mutation" if a change was made to the genome, and "experimental manipulation" if it was made to the niche. Generally we may call such perturbations *ablations*. Equivalently, a philosopher might talk of *counterfactuals*, i.e., what would happen if such-and-such an element of a system were missing. This is what is done in the high-throughput molecular biology laboratory. Large numbers of ablations are produced independently and in pairwise combination, allowing the definition of genetic interaction maps, defining the underlying information-processing structure of the organism. If we make enough independent recordable ablations, we can operationally define all the informational elements comprising the organism. Notice how ablations define the information that matters, that which confers real meaning -life or death - to the organism. Just as importantly, the procedure defines the ablations that do not matter.

If we can do enough experiments (i.e, with enough independent ablations) we can achieve full definition of the organism as an informationally closed entity. In other words, if we could continue obtaining ablations, we would reach a point where we would get no new ones. Any suggestion to the contrary would be to posit the inability to obtain an ablation, and this objection would be self-defeating, since if an ablation could not occur, it could not have an affect on the entity. Similarly, any objection as to the ability to define the niche takes the objector beyond the agreed definition of the niche in question.

We suggest the same process could allow the visualisation in principle of the dependency networks for the use of DOs have the appearance of genomic information; indeed, we could ground our position in the following example, in which we obtain a definite genetic interaction map for a known DO.

Let us consider a DO which is an actual recorded DNA sequence, such as the yeast genome obtained by DNA sequencing methodology, currently found within a digital library such as GenBank [3]. We could in principle synthesise DNA from this digital library information [4], insert it into a yeast cell lacking its own DNA, and use this cell to inoculate a culture within a growth chamber (its niche). If this culture performs its usual behaviour within its niche, it verifies the authenticity of the digital information. The test is straightforward: we do not need to know what behaviour to look for, we just compete the yeast (culture) containing the synthesised DNA with the wild or natural yeast (culture).

We could obtain a genetic interaction map of this DO, by performing ablations on the digital sequence of the object as well as on its digital environment, to figure out the boundaries of its Significant Properties (SP). Unimportant environment information disappears in this process, but environment information that matters - Significant Environment Information (SEI) - crystallise out in the interaction map, especially that information that influences the SP of the DO.

We could then perform the same test for functionality as above, and get the same map as we obtained using ablations in the natural DNA. This conceptual procedure would tell us that the DO maintains its significant properties; but its utility here is not in merely confirming the functional content of the information. Instead, it forces the recognition that the information in the DO, if it is to have utility at all, is exactly that prescriptive information that is contained in the natural DNA; the DO and the information in the genome are one and the same thing, by operational definition.

Thus, at least for yeast, the DO corresponding to the genome can be rendered into a useful map by ablation. In this case the recognition is made clear by employing the known identity between a DO and an organism's DNA sequence. Benefit may be gained by application of the heuristic. We *know* that every DO has a definite niche in its usage dependencies, just like an organism; we *know* that every DO comprises prescriptive information, SP, SEI and other bits in the DO and the environment, just like the genome of an organism; we *know* that just like in the genome, some of that information is historical nonsense, the "other bits", while other information (SP and the SEI) is crucially important. Crucially, this depends, on where we decide to throw the boundaries of the niche, which this perspective forces us to be clear about.

We analysed the kinds of information extractable from a DO and its environment to improve its chances of being useful in the long term. By this perspective we came to the conclusion that it is possible to extend the SP framework beyond the DO to its environment. This is the SEI [5] for a DO, defined as all the information needed, based on a particular purpose being addressed, to make use of it. Thereby SEI is a broad super-set of the existing SPs from where we adopt the concept of intended purpose, but extended to the whole DO environment and not just for the DO's intrinsic properties. See Fig.1.

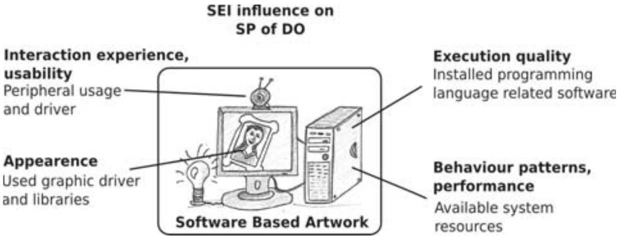

SEI influence on
SP of DO
Interaction experience, usability
Peripheral usage and driver
Appearence
Used graphic driver and libraries
Software Based Artwork
Execution quality
Installed programming language related software
Behaviour patterns, performance
Available system resources

Figure 1. SEI influences SP

Further we exemplified the above finding on a real-life software-based art preservation scenario using PERICLES Extraction Tool [5], a tool to extract SEI from the DE of a DO in a sheer curation [6] scenario, to improve the DOs reuse and the preservation of its SP that are influenced by the SEI. Sheer curation is a parallel to DE where organisms cannot be observed reliably outside of their niches, this resulting in an unavoidable loss of important information. To map their connectedness, a software agent observed and collected information about interactions between the DO and its immediate surroundings. By observing such interactions one can obtain a series of observations for further analysis and recognise functional dependencies. Such information cannot be reliably reconstructed after the DO is archived. It has to be extracted from the "live" system when the user is present, and preserved together with the DO.

Our theoretical model is visualised with the aid of this example on our corresponding poster.

3. ACKNOWLEDGMENTS

This work was supported by the European Commission Seventh Framework Programme under Grant Agreement Number FP7-601138 PERICLES.

4. REFERENCES

1. http://pericles-project.eu/
2. Fernando, C., Kampis, G., and Szathmáry, E. 2011. Evolvability of natural and artificial systems. In *Proceedings of the European Future Technologies Conference and Exhibition.*
3. GenBank ®: http://www.ncbi.nlm.nih.gov/genbank/
4. http://www.ncbi.nlm.nih.gov/genome/15
5. Corubolo, F., Eggers, A.G., Hasan, A., Hedges, M., Waddington, S., and Ludwig, J. 2014. A pragmatic approach to signifcant environment information collection to support object reuse, in IPRES 2014 proceedings.
6. http://alimanfoo.wordpress.com/2007/06/27/zoological-case-studies-in-digital-curation-dcc-scarp-imagestore/

A Biological Perspective on Digital Preservation

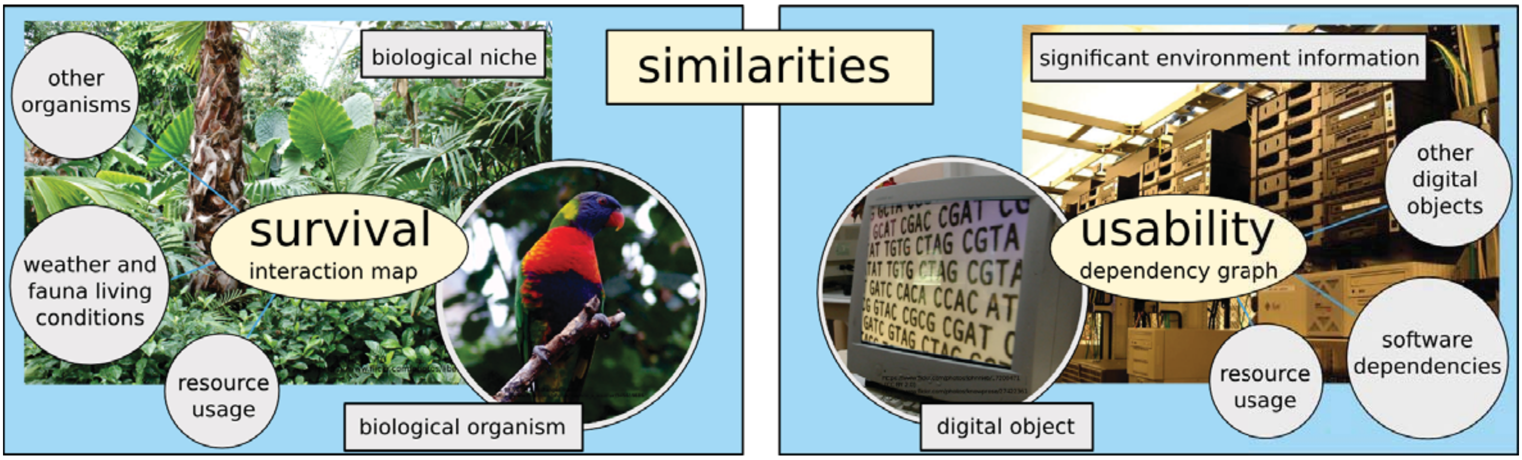
A transfer of Biological Ecosystem methodologies into the digital world and its benefits for Digital Preservation

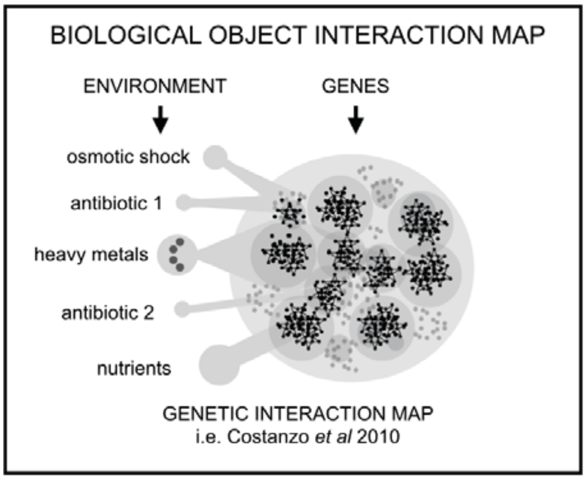
Successful preservation of Digital Objects (DOs) ultimately demands a solid theoretical framework. Such a framework with a high degree of generality emerges by treating DOs as containers of functional genetic information, exactly as in the genomes of organisms. We observe that functionality links survival in organisms and utility in DOs. In both cases, functional information is identifiable in principle by the consequence of its ablation.

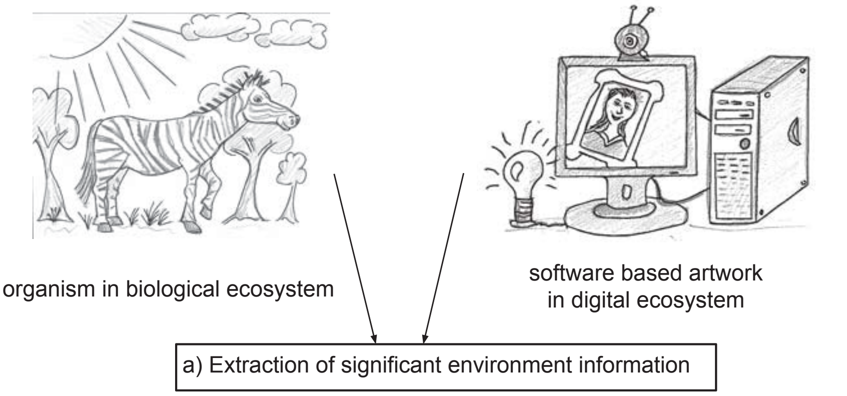
In molecular biology, genetic ablations (mutations) and environmental ablations (experimental manipulations) are used to construct interaction maps fully representing organismic activity. The equivalent of such interaction maps are dependency graphs for the use of DOs within their Digital Environment. These graphs are extracted live (sheer curation), and weighted based on the significance of the environment entities for the regarded DO uses.

The Biological Ecosystem

The Digital Ecosystem



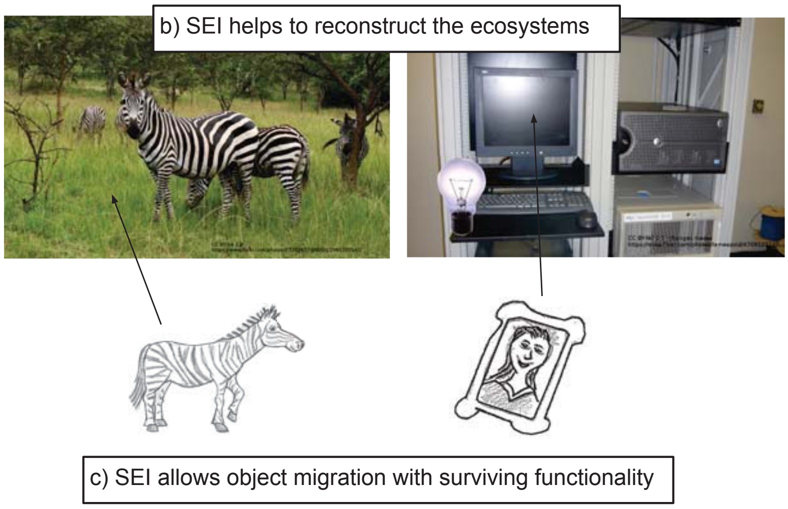




organism in biological ecosystem

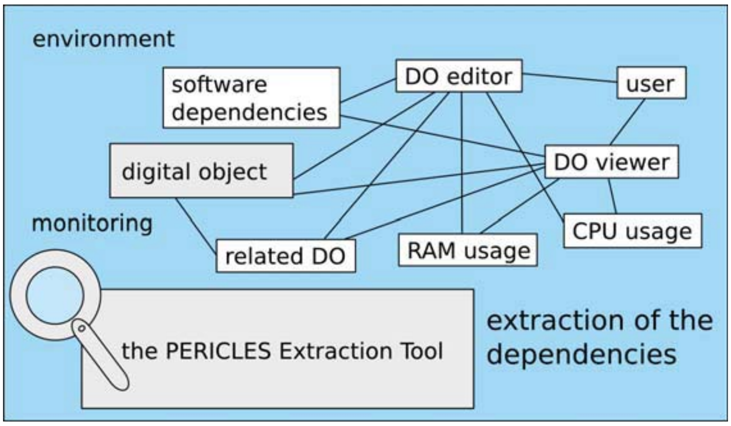
software based artwork in digital ecosystem

a) Extraction of significant environment information



b) SEI helps to reconstruct the ecosystems

c) SEI allows object migration with surviving functionality



environment

software dependencies

DO editor

user

digital object

DO viewer

monitoring

related DO

RAM usage

CPU usage

the PERICLES Extraction Tool

extraction of the dependencies

A biological view on digital preservation brings benefits for both disciplines

Validity of the biological perspective on DOs has been illustrated in the case of the yeast genome. The genome is both a biological object and a DO in its sequenced shape. They are united by a common theoretical framework based on functionality, and operationally defined by fitness in the face of analysis by mutations.

Generalised evolutionary perspective affords advantages to digital preservation in form of a better view of SEI than was hitherto possible.

The long term preservation perspective affords advantages to generalised evolution by supplying an example of a non-organismic evolutionary entity, by virtue of the fact that it is digital, allowing structural analysis. This is likely to be the first of many examples.

Pericles
FP7 Digital Preservation

This project has received funding from the European Union's Seventh Framework Programme for research, technological development and demonstration under grant agreement no FP7- 601138 PERICLES.

Legal Aspects for Digital Preservation Domain

Barbara Kolany-Raiser,
Institute for Information,
Telecommunication and Media Law
(ITM), WWU
Muenster, Germany
barbara.kolany@uni-muenster.de

Marzieh Bakhshandeh,
José Borbinha,
Instituto Superior Técnico,
Universidade de Lisboa and
Information Systems Group, INESC-ID
Lisboa, Portugal
{marzieh.bakhshandeh,jlb}@ist.utl.pt

Silviya Yankova
Institute for Information,
Telecommunication and Media Law
(ITM), WWU
Muenster, Germany
silviya.yankova@uni-muenster.de

ABSTRACT

Long term digital preservation serves the preservation of data substance and operability, so that future users are enabled to use stored data and rerun the preserved processes to gain the stored information. Furthermore, Law is becoming an essential application domain for technology developments. In case copyright protected data has to be digitally preserved, every process of a digital preservation system may violate this right, when the rightholder who has the exclusive rights did not grant the relevant rights of use. This paper shows a Legal Ontology that provides a hierarchical overview of how legal constraints and obligations (e.g. IP rights and licensing issues) could be implemented in an automated process of a DP system. In simply terms, difficulties with legal taxonomies may arises when the creators and the users don't share the same perspective. This would be the case when the creators of the taxonomy are lawyers and the users not. Legal taxonomies for digital preservation can be represented with ontologies which are an explicit account of a shared understanding in any domain. Through the use of ontologies the communication can be improved, which, in turn, can give rise to greater reuse, sharing, transparency, and inter-operability. Every DP activity must ensure the authenticity and legitimacy of the performed actions and processes. Hence to validate the correctness of our legal ontology we used a set of competency questions defined in a specific case study. The goal is to obtain a clearer taxonomical view of the necessary legal knowledge that will address the concerns of industrial use-case DP stakeholders. Therefore, we recommend using the Legal Ontology for the DP domain, in order to integrate different legal perspectives and perform reasoning and inference over legal knowledge and information.

Digital Preservation (DP) does exist for a long time and is an ongoing challenge for information society. Heretofore the main focal point has been on the preservation of static digital objects and artifacts. The TIMBUS EU Project[1] uses this fundus of information to develop solutions which enables the preservation of interactive media, dynamic digital objects, and entire business processes and services.

The description of whole processes including all their inter-dependencies, essential components and their configurations is a complex task. The aim is to re-deploy the systems in the future and to do this in a way which allows interaction with them. To ensure the authenticity and legitimacy of the performed actions and processes is an essential part of every DP activity. In order to deal with different legal perspectives and concerns, the ontological approach can help to organize legal information and requirements– making it a pivotal element of any DP system. It is obvious that legal issues and obligation in the DP have to be addressed. Rights can be infringed by almost any process of a DP system. Besides, there are other legal requirements involved, e.g. contracting issues and licensing. In order to reach our aim of creating a common understanding of the meaning of legal concepts and terms, ontologies can help to mitigate the risk of misinterpretation, especially in the field of in legal applications, by giving contextual explanation and precise legal information. The importance of this technology is evidenced by the growing use of ontologies in a variety of application areas [1] ,[2], and[3]. Also, by their role on the Legal areas as observed in [4],[5], [6], [7], [8] and [9].
In the following, we first want to outline the importance of legal aspects for digital preservation and then briefly introduce the concept of Legal Ontology Engineering in the domain of DP. After emphasizing the drawbacks of these works, we present our methodology to address these shortcomings. Then we focus on showing the innovation and advantages of our developed Legal Ontology on the basis of a recent case-study in e-Health. Finally, we point out specific validation steps taken to evaluate our work and sum up our contributions to complete the paper.

The preservation of digital objects and the reuse of them in the future are influenced by legal requirements. This has effects on all aspects of the preservation challenge: business constraints, process descriptions, computational environments and their mutual dependencies, digital assets that are produced and consumed by the processes, roles of individuals and organisations, and dependencies on third-party products and services. These requirements are established in European Directives as well as national laws or regulations which have a large impact on how Digital Preservation can be carried out. Preservation actions might have implications on intellectual property rights or data protection. To find out what legal requirements must be fulfilled the first question has to be if data that should be digitally preserved is protected. When data is copyright protected it is important to know what the terms of the license contract determine regarding the right of use. Even when already existing

[1] http://timbusproject.net/

digital objects have to be moved from one folder to another within the digital preservation system copyright might be affected. The storage of copyright protected data can potentially infringe the copyright-holder’s exclusive rights when the act of reproduction is not allowed. The main goal of Digital Preservation is to keep important information available for the future. Hence it is important to take a closer look at the activities that might be possible and necessary activities in the digital preservation system. Processes like migration or emulation are very important to keep the stored information safe for the future. The admissibility of these actions depends on the terms of the license contract. Therefore, if the existing license contracts do not allow such actions, amendments might be necessary. The setting-up and optimizing of IT contracts need to compensate the various interests of the stakeholders who are involved in the preservation efforts. Digital preservation actions might not only cause potential violation of copyrights, but also might infringe data protection law through the storage of personal data. Questions arise like: does a prior and valid consent of the data subjects exist; does a legal permission to store the data exist; where will the data be stored. Data protection requirements differ in the various EU-Member states. To comply with the legal requirement regarding data protection law it is even more difficult when storage is planned to be used outside the EU, e.g. in the United States. Furthermore, legal requirements like the fulfillment of legal obligations to preserve certain data can be a driver for digital preservation. All enterprises have to retain and preserve data, the so called non-sector specific preservation obligations established in e.g. tax law or commercial law. Besides, there might also exist corresponding additional obligations, the so called sector-specific obligations. The law identifies what must be preserved, for how long and for what purpose. The clearer you have in mind what the legal requirements are you have to fulfill, the better you can think of strategies to avoid potential infringement.

Keywords

Digital Preservation, Ontology, Legal Ontology, Legal taxonomies

http://timbusproject.net

Legal Aspects for Digital Preservation Domain

Barbara Kolany-Raiser[1], Marzieh Bakhshandeh[2], José Borbinha[2], Silviya Yankova [1]

[1] Institute for Information, Telecommunication and Media Law (ITM), WWU Muenster, Germany

[2] Instituto Superior Técnico, Universidade de Lisboa and Information Systems Group, INESC-ID Lisboan, Portugal

We propose a legal ontology for the digital preservation domain.

Ontologies describe a domain model by associating meaning to its terms and relations. The importance of this technic is evidenced by the growing use of ontologies in a diversity of application areas.

This unifying **Legal Ontology** is intended to function as a lingua-franca to facilitate the translation and mapping between different perspectives, as well as reasoning and inference over legal information in the domain of digital preservation. Next, the legal ontology was validated by a set of competency questions through two specific case study. This validation was processed with reasoning methods.

Work in progress is focusing on the application of this approach to multiple scenarios...

Law is becoming an essential application domain for technology developments. In case copyright protected data has to be digitally preserved, every process of a digital preservation system may violate this right, when the rights holder who has the exclusive rights did not grant the relevant rights of use.

We developed a **Legal Ontology** that provides a hierarchical overview of how legal constraints and obligations (e.g. IP rights and licensing issues) could be implemented in an automated process of a digital preservation system.

In simply terms, difficulties with legal taxonomies may arise when the creators and the users don't share the same perspective. This would be the case when the creators of the taxonomy are lawyers and the users not. Legal taxonomies for digital preservation can be represented with ontologies which are an explicit account of a shared understanding in any domain.

Through the use of ontologies the communication can be improved, which, in turn, can give rise to greater reuse, sharing, transparency, and inter-operability.
Every digital preservation activity must ensure the authenticity and legitimacy of the performed actions and processes. Hence to validate the correctness of our legal ontology we used a set of competency questions defined in a specific case study. The goal is to obtain a clearer taxonomical view of the necessary legal knowledge that will address the concerns of industrial use-case digital preservation stakeholders.

Therefore, we recommend using the **Legal Ontology** for the digital preservation domain, in order to integrate different legal perspectives and perform reasoning and inference over legal knowledge and information.

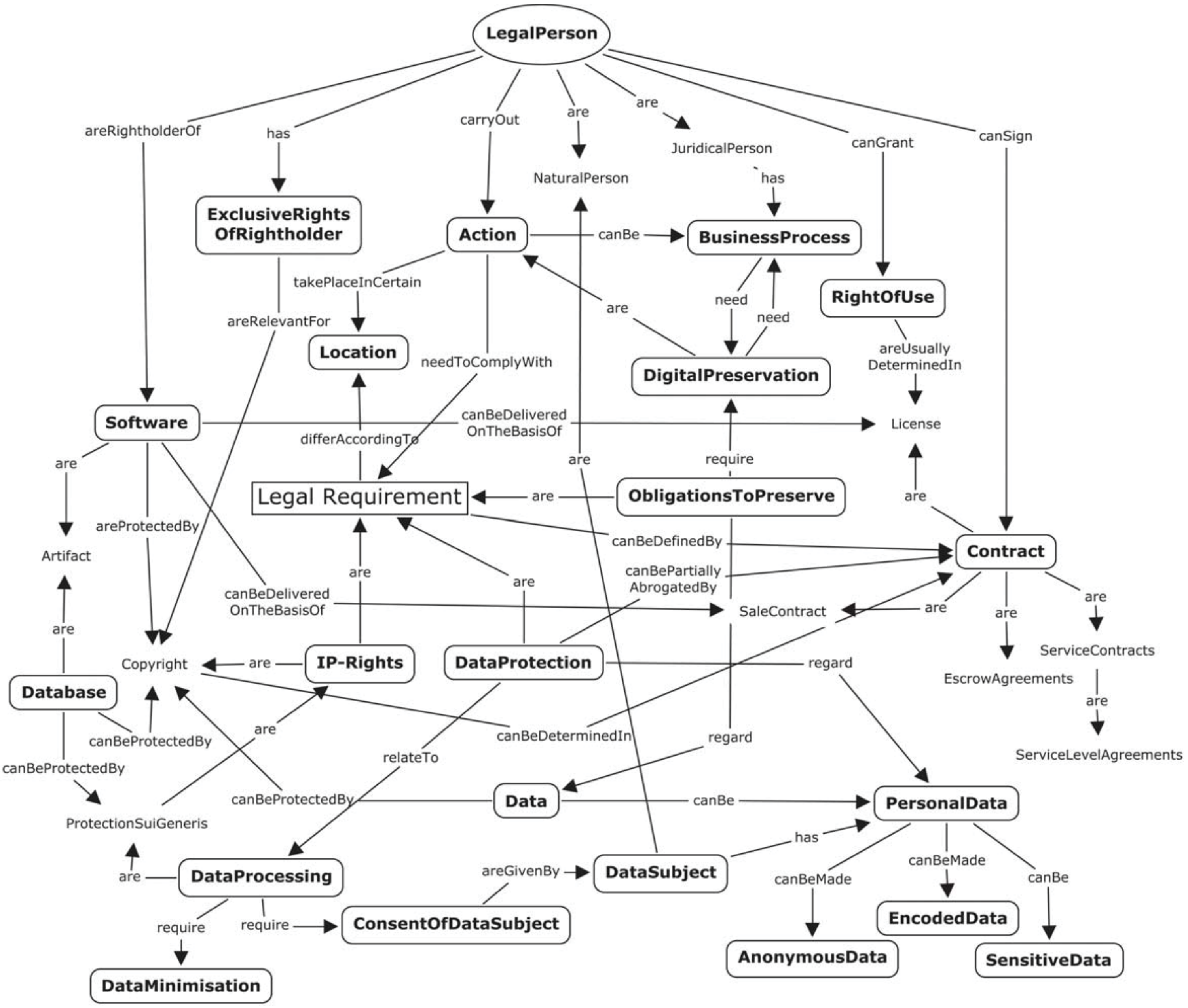
Ontology is an explicit formal specification of the terms in a domain and relations among them – basically similar to a taxonomical representation of a class hierarchy in a given domain. Ontologies describe structure and hierarchy.

Ontologies play an important role in knowledge sharing in the field of knowledge representation and reasoning The ontology building process is a craft, rather than an engineering activity. The steps include:

(1) identification of the concepts and concept hierarchy
(2) identification of the disjoint concepts
(3) modeling composition
(4) addition of all the relationships between concepts
(5) identification of definitions
(6) addition of annotations
(7) and refinement of the ontology through various iterations of the above steps.

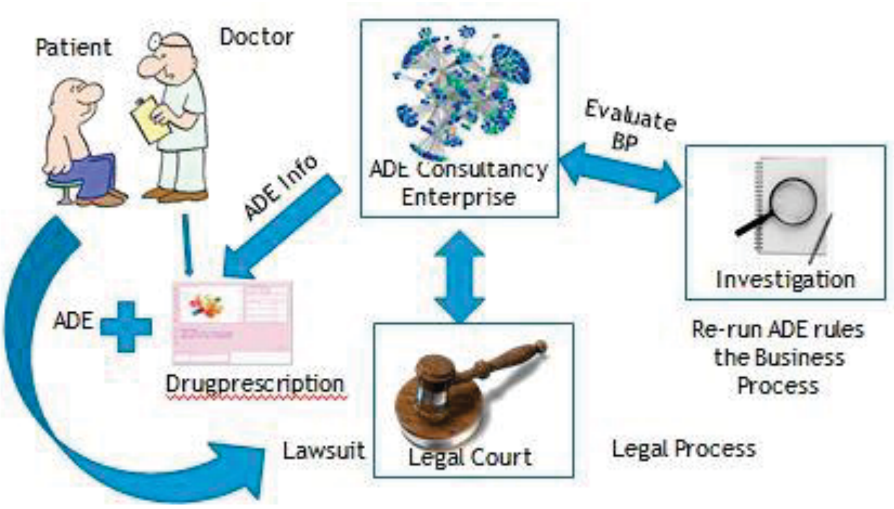
Most ontology building methods propose iterative approaches in order to allow formalization to be accomplished progressively.

In this work, we followed an iterative approach by using conceptual maps as a “bridge” between the legal taxonomy and the formal specification. For the first phase, the concepts and their relationship were drawn in a Conceptual Map model which depicts a representation of the conceptual map used to develop our Legal Ontology. We can see a conceptual map of the legal perspective. In this description the concepts are written in bold and the relationships are in italic.



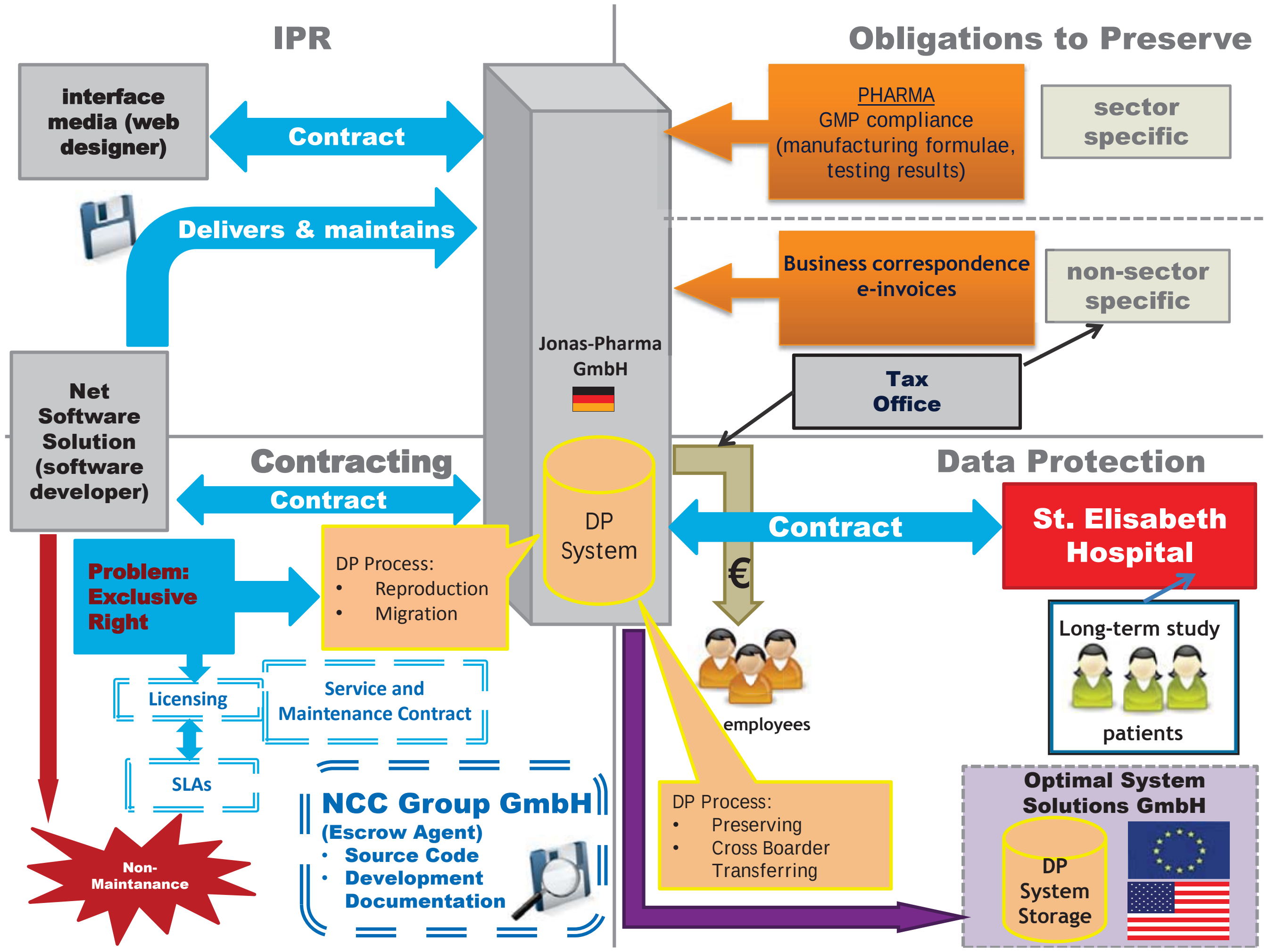
A case-study of an e-Health scenario

It is concerned with addressing the ADR problem by providing a web-based solution for discovery and search of ADE (Adverse Drug Event) rules used by doctors and pharmacists for prescribing drugs.



A case-study in pharma

Jonas-Pharma GmbH is a Pharmaceutical Company with its headquarters in Cologne and enters into a License Contract with a Software Development Company, Net Software Solution, in order to use the software Iris created by that Software Development Company. The Jonas-Pharma GmbH wants to digitally preserve the relevant data of their business processes including the software Iris. Consequently, the necessary rights of use must be granted in the License Contract. The rights of re production and migration and alteration are essential for digital preservation. In the given scenario, the necessary rights are not explicitly included. Consequently, an amendment agreement is required granting the necessary rights for digital preservation. The software Iris is copyright protected. Copyright belongs to the IP-Rights.



IPR

Obligations to Preserve

interface media (web designer)

Contract

Delivers & maintains

PHARMA
GMP compliance (manufacturing formulae, testing results)

sector specific

Business correspondence e-invoices

non-sector specific

Jonas-Pharma GmbH

Tax Office

Net Software Solution (software developer)

Contracting

Contract

DP System

Data Protection

Contract

St. Elisabeth Hospital

€

Problem: Exclusive Right

DP Process:
- Reproduction
- Migration

employees

Long-term study

patients

Licensing

Service and Maintenance Contract

SLAs

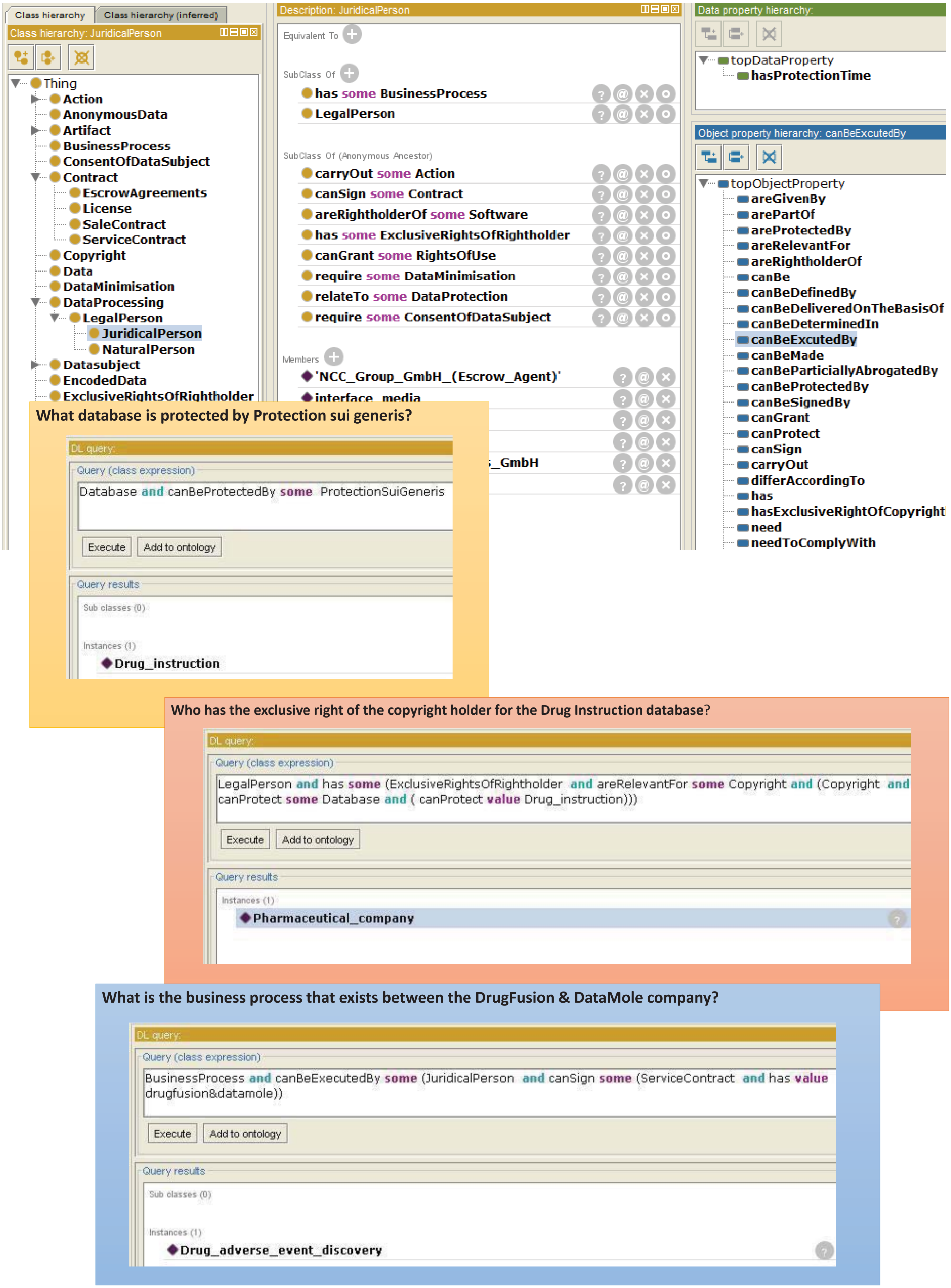
NCC Group GmbH
(Escrow Agent)
- Source Code
- Development Documentation

DP Process:
- Preserving
- Cross Boarder Transferring

Optimal System Solutions GmbH

DP System Storage

Non-Maintanance



Class hierarchy
Class hierarchy (inferred)
Class hierarchy: JuridicalPerson

- Thing
  - Action
  - AnonymousData
  - Artifact
  - BusinessProcess
  - ConsentOfDataSubject
  - Contract
    - EscrowAgreements
    - License
    - SaleContract
    - ServiceContract
  - Copyright
  - Data
  - DataMinimisation
  - DataProcessing
    - LegalPerson
      - JuridicalPerson
      - NaturalPerson
  - Datasubject
  - EncodedData
  - ExclusiveRightsOfRightholder

Description: JuridicalPerson
Equivalent To
SubClass Of

- has some BusinessProcess
- LegalPerson

SubClass Of (Anonymous Ancestor)

- carryOut some Action
- canSign some Contract
- areRightholderOf some Software
- has some ExclusiveRightsOfRightholder
- canGrant some RightsOfUse
- require some DataMinimisation
- relateTo some DataProtection
- require some ConsentOfDataSubject

- 'NCC_Group_GmbH_(Escrow_Agent)'
- interface_media
- _GmbH

- topDataProperty
  - hasProtectionTime

- topObjectProperty
  - areGivenBy
  - arePartOf
  - areProtectedBy
  - areRelevantFor
  - areRightholderOf
  - canBe
  - canBeDefinedBy
  - canBeDeliveredOnTheBasisOf
  - canBeDeterminedIn
  - canBeExcutedBy
  - canBeMade
  - canBeParticiallyAbrogatedBy
  - canBeProtectedBy
  - canBeSignedBy
  - canGrant
  - canProtect
  - canSign
  - carryOut
  - differAccordingTo
  - has
  - hasExclusiveRightOfCopyright
  - need
  - needToComplyWith

What database is protected by Protection sui generis?

```
Database and canBeProtectedBy some  ProtectionSuiGeneris
```

Sub classes (0)

Instances (1)

- Drug_instruction

Who has the exclusive right of the copyright holder for the Drug Instruction database?

```
LegalPerson and has some (ExclusiveRightsOfRightholder  and areRelevantFor some Copyright and (Copyright  and
canProtect some Database and ( canProtect value Drug_instruction)))
```

Instances (1)

- Pharmaceutical_company

What is the business process that exists between the DrugFusion & DataMole company?

```
BusinessProcess and canBeExecutedBy some (JuridicalPerson  and canSign some (ServiceContract  and has value
drugfusion&datamole))
```

Sub classes (0)

Instances (1)

- Drug_adverse_event_discovery

This work was supported by national funds through FCT (Fundação para a Ciência e a Tecnologia), under projects Pest-OE/EEI/LA0021/2013, TAGS PTDC/EIA-EIA/112283/2009, and by the project TIMBUS, co-funded by the European Commission under the 7th Framework Programme for research and technological development and demonstration activities (FP7/2007-2013) under grant agreement no. 269940.

Demonstrating a Digital Curation Workflow using the BitCurator Environment

Christopher A. Lee
School of Information and Library Science
University of North Carolina
216 Lenoir Drive, CB #3360
1-(919)-966-3598
callee@ils.unc.edu

ABSTRACT

This demonstration will highlight several key steps in a digital curation workflow that incorporates digital forensics tools and methods. Using the open-source BitCurator environment, I will demonstrate several discrete tasks, how they can feed into each other, and considerations related to incorporating them into a larger set of curation practices within collecting institutions. A strong emphasis will be placed on features of the software that have been added or enhanced over the past year, including mounting and exporting of files from forensically packaged disk images, identification of duplicate files, generation of PREMIS metadata and initial steps toward redaction of potentially sensitive information.

Categories and Subject Descriptors

H.3.7 [**Information Storage and Retrieval**]: Digital Libraries – *collection, dissemination, systems issues.*

General Terms

Provenance, Data Triage, Digital Forensics.

Keywords

Forensics, preservation, DFXML, metadata, privacy, collections, acquisition

1. BITCURATOR PROJECT

The BitCurator Project, a collaborative effort led by the School of Information and Library Science at the University of North Carolina at Chapel Hill and Maryland Institute for Technology in the Humanities at the University of Maryland, is addressing two fundamental needs and opportunities for collecting institutions: (1) integrating digital forensics tools and methods into the workflows and collection management environments of libraries, archives and museums and (2) supporting properly mediated public access to forensically acquired data [4].

2. BITCURATOR ENVIRONMENT

We are developing and disseminating a suite of open source tools. These tools are being developed and tested in a Linux environment; the software on which they depend can readily be compiled for Windows environments (and in most cases are currently distributed as both source code and Windows binaries). We intend the majority of the development for BitCurator to support cross-platform use of the software. We are freely disseminating the software under an open source (GPL, Version 3) license. BitCurator provides users with two primary paths to integrate digital forensics tools and techniques into archival and library workflows.

First, the BitCurator software can be run as a ready-to-run Linux environment that can be used either as a virtual machine (VM) or installed as a host operating system. This environment is customized to provide users with graphic user interface (GUI)-based scripts that provide simplified access to common functions associated with handling media, including facilities to prevent inadvertent write-enabled mounting (software write-blocking).

Second, the BitCurator software can be run as a set of individual software tools, packages, support scripts, and documentation to reproduce full or partial functionality of the ready-to-run BitCurator environment. These include a software metapackage (.deb) file that replicates the software dependency tree on which software sources built for BitCurator rely; a set of software sources and supporting environmental scripts developed by the BitCurator team and made publicly available at via our GitHub repository (links at http://wiki.bitcurator.net); and all other third-party open source digital forensics software included in the BitCurator environment.

3. DEMONSTRATED TOOLS AND FEATURES

Tools that BitCurator is incorporating include Guymager, a program for capturing disk images; bulk extractor, for extracting features of interest from disk images (including private and individually identifying information); fiwalk, for generating Digital Forensics XML (DFXML) output describing filesystem hierarchies contained on disk images; The Sleuth Kit (TSK), for viewing, identifying and extraction information from disk images; Nautilus scripts to automate the actions of command-line forensics utilities through the Ubuntu desktop browser; and sdhash, a fuzzing hashing application that can find partial matches between similar files. For further information about several of these tools, see [1,2,3,5].

This demonstration place significant emphasis on features of the software that have been added or enhanced over the past year, including mounting and exporting of files from forensically packaged disk images, identification of duplicate files, generation of PREMIS metadata and initial steps toward redaction of potentially sensitive information. Other supported features that will be illustrated in the demonstration include mounting media as read-only, creating disk images, using Nautilus scripts to perform batch activities, generation of Digital Forensics XML (DFXML), generation of customized reports, and identification of sensitive data within data.

4. ACKNOWLEDGMENTS

BitCurator development has been supported by the Andrew W. Mellon Foundation. Members of the BitCurator team are Alexandra Chassanoff, Matthew Kirschenbaum, Christopher (Cal) Lee, Sunitha Misra, Porter Olsen, and Kam Woods. Members of two advisory boards have made valuable contributions: the Development Advisory Group (DAG) and Professional Experts Panel (PEP).

5. REFERENCES

[1] Cohen, M., Garfinkel, S., and Schatz, B. 2009. Extending the Advanced Forensic Format to Accommodate Multiple Data Sources, Logical Evidence, Arbitrary Information and Forensic Workflow. *Digital Investigation* 6 (2009), S57-S68.

[2] Garfinkel, S. Digital Forensics XML and the DFXML Toolset. *Digital Investigation* 8 (2012), 161-174.

[3] Garfinkel, S.L. Providing Cryptographic Security and Evidentiary Chain-of-Custody with the Advanced Forensic Format, Library, and Tools. *International Journal of Digital Crime and Forensics* 1, 1 (2009), 1-28;

[4] Lee, C.A., Kirschenbaum, M.G., Chassanoff, A., Olsen, P., and Woods, K. BitCurator: Tools and Techniques for Digital Forensics in Collecting Institutions. *D-Lib Magazine* 18, 5/6 (May/June 2012).

[5] Roussev, V. An Evaluation of Forensic Similarity Hashes. *Digital Investigation* 8 (2011), S34-S41.

Curation Cost Exchange platform

Luis Faria
KEEP SOLUTIONS
Braga, Portugal
+351 253066735
lfaria@keep.pt

Neil Grindley
JISC
London, UK
+44 (0)203 006 6059
n.grindley@jisc.ac.uk

ABSTRACT

This demonstration proposal describes the Curation Cost Exchange platform (CCEx), a web application that allows organizations to introduce, analyse, share and compare the cost of their digital curation activities. It is also a central hub for digital curation costing related information; and is a social platform that brings together organizations with the same problems and allows sharing of experiences, good practices and know-how. The CCEx is an output of the 4C Project (a Collaboration Clarify the Costs of Curation) and the relationship of the CCEx to other 4C Project outputs will also be briefly described.

General Terms

Communities, strategic environment, digital preservation marketplace, case studies and best practice.

Keywords

Curation, Cost, Cost analysis, Economy, Curation activities, Cost analysis, Cost comparison, social, Cost information, Cost model.

1. INTRODUCTION

A lot of excellent and detailed work has been carried out over the last decade to develop and refine cost models for digital curation and it is now possible to make an assessment of those methods and to design a new approach for tackling this very complex problem.

Improved clarity about the costs of digital curation supports tactical and strategic decision-making within an organisation and will improve the efficiency of digital asset management. The current problem is that there is no authoritative cost model that can be generically employed and there is little by way of comparative data that organisations can benchmark themselves against. This results in individualistic methods and no clear path to understanding what the typical or acceptable costs are for digital curation activities. By enabling organizations to share and compare the costs of curation activities with each other, benchmark costs for various curation activities can emerge and organizations can better assess how they should spend their budgets and plan their investments. Knowing what similar organizations have spent on curation activities (and why they have prioritised that spend) is a valuable insight.

Organizations organize their costs in very particular ways and this makes it hard to directly compare costs. To solve this, the 4C project devised a framework to map costs into a set of activities, capital procurements and labour roles, enabling organizations to compare costs in these categories.

This framework is used within a web application, the Curation Cost Exchange Platform, which enables an organization to submit their costs online and then compare them with other organizations.

2. Framework of comparable costs

The framework defines a set of cost categories into which the curation costs of an organization can be mapped, allowing different organizations to directly compare within those categories.

The primary mapping is done to an OAIS [1] based set of activities: production, ingest, archival storage and access. This mapping allows organizations with activity based accounting to easily map their costs into a set of categories that plainly divide curation concepts.

A secondary mapping is based on financial accounting, dividing costs into capital procurements: hardware, software, external or third party services; into labour roles: producer, IT-developer, support/operations, records manager and manager; and into overhead. This secondary mapping is closer to financial accounting which is further away from the curation concepts but is closer to the usual accounting practice in organizations.

3. Curation Cost Exchange platform

The Curation Cost Exchange platform (CCEx) is a web application that allows users to submit information about the curation costs in their organizations, map into the categories defined in the framework, and analyse the resulting self-assessment, group and peer-to-peer comparison.

3.1 Submission template

A web based submission template allows users to define a profile of their organization and its collections of assets, describing the characteristics that might affect costs. This information enables matches with similar organizations against which the cost comparison is more appropriate.

The web submission form then requires the input of a list of cost units, which refer to the costs on the organizations own structure. Each cost unit can be mapped to the concepts introduced on the framework of comparable costs by using percentage ratios. The mapping is validated so no overlaps exist on cost mapping.

Finally, the costs of each category are harmonized per data volume, providing costs per Gigabyte. These relative costs allow direct numeric comparisons of costs between organizations on the categories defined by the framework of comparable costs.

The user can now analyse the result as their own self-assessment of costs, or compare their costs with other organizations, either as a group or peer-to-peer.

3.2 Self-assessment

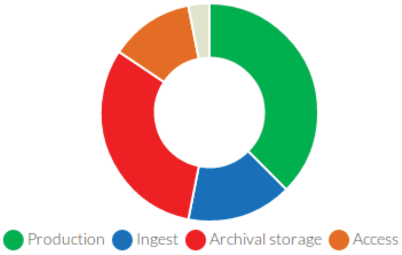


Figure 1. Example comparison of the budget spent on different curation activities within an organization.

The analysis of organizational costs mapped into the framework categories allows a level of self-assessment of how the budget is being spent. The web application shows the comparison of the mapping into each of the categories, for both activity and financial accounting. The result, as shown in Figure 1, is a doughnut chart that compares the categories by cost.

3.3 Group comparison

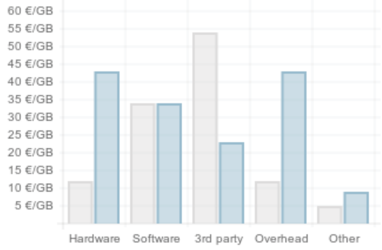


Figure 2. Example comparison of budget spent in capital procurements with the average of all other organizations.

The web application allows comparison of the costs, mapped into the framework categories, with the average of all (or a subset of) organizations that have also submitted costs into the platform[1]. Figure 2 shows an example of the comparison of an organization's costs mapped into capital procurements with all other organizations.

Different types of organizations might have costs that are not comparable with each other, like comparing costs of national libraries with small or medium enterprises, or comparing costs of organizations that have mainly audio-visual material with others that only have text documents. This and other cost determinants are used to allow comparison filtering, ensuring that the organization costs are compared against similar ones, which provides a more valuable and trusted cost reference.

The statistical analysis of the submitted costs and the organization and collection characteristics allows the definition on new cost determinants and improvement of the filters that allow valuable group comparisons.

3.4 Peer-to-peer comparison

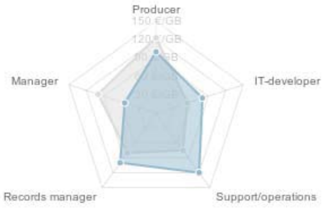


Figure 3. Example comparison of budget spent in different labour roles between two organizations.

The web application also allows peer-to-peer comparison between organizations with similar characteristics, to find out discrepancies. Figure 3 shows an example comparison of the costs mapped into labour roles between two organizations. This comparison is only possible if at least one of the organizations allows peer-to-peer comparison, although it can maintain anonymity.

A communication channel between the two organizations can be requested for both organizations to get in contact and share experiences and best practices.

4. Conclusion

The Curation Costs Exchange is one of the core deliverables of the 4C Project and is an ambitious attempt to try and tackle a self-perpetuating problem. In the past, organisations wishing to understand the costs of curation have discovered that cost models designed by others are difficult to use and that there is very little comparative data that is publicly available to benchmark activity against. This has forced them to devise their own calculation methods and has not incentivised them to share their costs data with others. The point of the CCEx and the 4C work more generally is to try and harmonise practice and encourage data sharing.

One of the 4C Project principles is to be 'open and social' and it is this collaborative approach that we believe will ultimately help the community to get a better grasp of the costs of curation. We will also be drawing up a Roadmap and an action agenda for post-project activity that will further define and support the need for future collaborative action and will also set out a sustainability path for the CCEx and other critical 4C outputs.

5. ACKNOWLEDGMENTS

The 4C project is funded by the European Commission as part of the FP7 ICT Cooperation programme, contract 600471.

6. REFERENCES

[1] Consultative Committee for Space Data Systems. Reference Model for an Open Archival Information System (OAIS). CCSDS 650.0-B-1. Washington, DC, 2002.

[1] Only organizations which have agreed to share costs are included in the cost comparison.

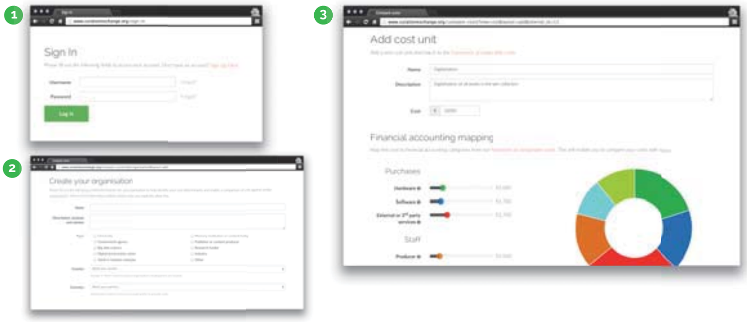
Curation Costs Exchange

Understanding and comparing digital curation costs to support smarter investments

Compare costs

Add your curation costs and see how you compare to others

1 Sign in ✓ | 2 Organisation profile ✓ | 3 Cost data sets ✓ | 4 Compare costs



Sign In

Username

Password

Log in

Create your organisation

Add cost unit

Name

Description

Cost

Financial accounting mapping

Purchases

Hardware

Software

External or 3rd party services

Staff

Producer

4 **Global comparison**

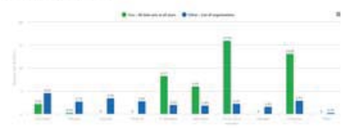
Filter by organisation type, asset types, data volume, staff size, and others.

Other organisations

List all organisations

● List all organisations	12
○ Collections with the same number of copies	6
○ Organisations of type Memory institution or content holder	7
○ Collections with around the same staff	3
○ Collections of scope: A collection	5

Financial accounting comparison



Peer comparison

Select to compare with organisations alike yours, sorted by similarity to your own profile.

KEEP SOLUTIONS contact

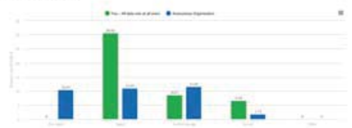
High similarity

KEEP SOLUTIONS is a **Digital preservation vendor, Small or medium enterprise** from **Portugal** with a digital curation staff of average **12 people** and a data volume of average **2 Terabytes**. It has a number of copies policy of average **2 replicas**.

KEEP SOLUTIONS is a Portuguese company that provides advanced services for managing and preserving digital information. The company initiated its activity in 2008 and ... read more

Compare with other peers

Activities comparison



Understand your costs

Understand how to assess your curation costs and how to make use of cost models to help you invest

Core cost concepts

How can I learn more about the key concepts in the costs and benefits of digital curation?

Economic sustainability reference model

How do I invest strategically to preserve data for the long term?

Indirect cost drivers

How do I get best value from my digital assets?

Quality and trustworthiness

Should I get certified - what are the costs and benefits of investing in a Trusted Digital Repository?

Model requirements specification

How do I develop my own cost model?

Summary of cost models

How can I get an overview of existing cost and benefit models?

OTHER FEATURES

Read more

Browse books, papers and articles to help you get started in curation costing.

Discuss and share

Share your experiences and read about challenges in other organisations like yours.

Find services

See a list of digital curation tools and service providers, and find out more about what they offer.

EUROPEAN RESEARCH PROJECT

This work has been developed under the 4C project, www.4cproject.eu, which has received funding from the European Union's Seventh Framework Programme for research, technological development and demonstration under grant agreement no. 600471.

The 4C project partners:

- Jisc
- The Royal Library of Denmark
- INESC-ID, Portugal
- Danish National Archives
- Deutsche Nationalbibliothek
- HATII, University of Glasgow
- University of Essex
- Keep Solutions
- Digital Preservation Coalition
- Secure Business Austria
- Digital Curation Centre, University of Edinburgh
- Data Archiving and Networked Services
- National Library of Estonia

TESTIMONIES

Armin Straube

Nestor - the German network for digital preservation

> "CCEx is an important tool for a best-practice network like nestor. It helps the digital preservation community to become more cost-efficient and professional and integrates well with our overall efforts."

Ron Dekker

NWO - Netherlands Organisation for Scientific Research

> "The CCEx is the platform to help funders realise the benefit of their investments. By being transparent about their costs and plugging them into this platform, projects can demonstrate that the taxpayer is getting value for money."

USAGE

From August 11 to September 25, 2014

Sessions

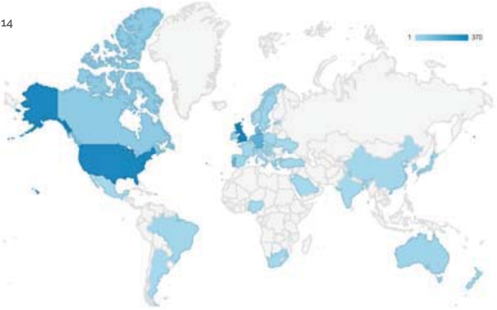
1,559

Pageviews

8,719

Avg. Session Duration

00:05:58



NETWORK

Organisations

Have pledged to share their curation costs

Authors

Luis Faria lfaria@keep.pt

Neil Grindley n.grindley@jisc.ac.uk

4C project team info@4cproject.eu

www.curationexchange.org

http://bit.ly/ipres14

Demonstration of an Integrated System for Platform-independent Description of Human-Machine Interactions

Oleg Stobbe, Klaus Rechert and Dirk von Suchodoletz
Albert-Ludwigs University Freiburg
Hermann-Herder Str. 10
79104 Freiburg i. B., Germany

1. MOTIVATION

When using emulation to render digital objects, a dedicated system environment is required. This environment typically consists of a set of software, i.e. an emulator, replicating the original hardware, operating system, hardware drivers, application as well as tools and utilities. Typically such technical meta-adata is modeled using a view-path. Configuration and operating knowledge, however, is also required and needs to be described and preserved to secure deterministic future environment and workflow replication.
One possible solution is to capture and replay human-machine interaction in an abstract way. A model for recording and capturing interactions between human users and an emulated machine has already been proposed [2]. With the integration of such a system in an Emulation-as-a-Service service model [1], usability has been improved significantly by integration the capturing and replay into EaaS workflows. This demo's purpose is to demonstrate the system's usability and utility for digital preservation tasks like automation, documentation and replication of interactive tasks.

2. ARCHITECTURE & IMPLEMENTATION

To capture and replay any user-interaction either directly with the running emulated system or with the emulator, an interaction workflow description (IWD) recorder is added to EaaS's emulation components. Emulation components abstract each emulator's individual complexity and provide unified interfaces for interaction with the emulated environment. In contrast to so-called macro-recorder, the IWD-system does not interact directly with the emulated operating system and thus, is platform independent and extensible to cover new, upcoming interaction paradigms.
The basic idea of IWD is to simulate a human user's behavior: before executing a single interaction, e.g. mouse movement, mouse click, keyboard input, the system needs to be in the appropriate state, i.e. providing a proper context for a certain action and a potential previous event has to be processed completely. More formally, a single event is described through a precondition, i.e. the system has to be in a specific, pre-defined state pc, an action a, and the expected outcome eo of the user-action $(ev_i :=< pc, a >_i \rightarrow eo_i)$. Both, pre- and postcondition of each interaction are verified by using the emulator's visual output and the emulator's internal machine state. Furthermore, we assume that each postcondition is also precondition for the next event.
In the current version synchronization is implemented using visual/graphical output only. Before executing the next action, the system waits for the screen to reach a state "similar enough" to the one, at the time of the action's recording. Pre- and post-conditions are rather simply automatically generated by fingerprinting the emulators output.

2.1 Recording Architecture

When using EaaS to instantiate a dedicated emulated system environment the user is able to simply enable recording of the session's interactions. Recording is performed on the server-side running as a background task independently of the way the user interacts with the emulator (e.g. using a web client or a dedicated VNC, RDP, etc.). An abstract *interceptor* interface allows to capture, filter and manipulate any message sent between the user's client and the emulation component. On the server-side, two worker threads analyze the user's stream of input events and the emulator's output. Specific sync-instructions and timestamps are used for synchronization of both streams. To support visual synchronization, the emulator's visual output is reconstructed

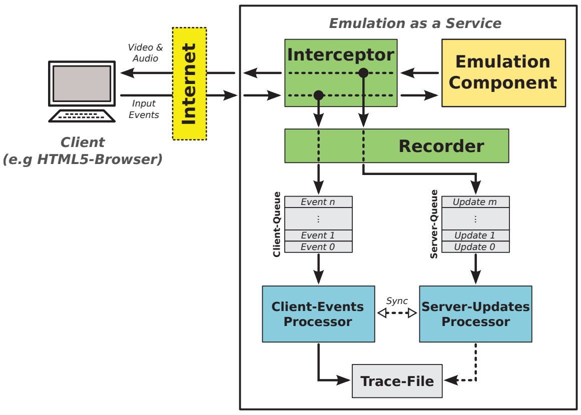


Figure 1: Recording of human-machine interactions using an EaaS setup.

and drawn on the server-side. For instance, all screen updates are processed before a mouse-click event. From this, a snapshot of $n \times m$ pixels surrounding the mouse cursor is written to the trace file. This also design provides some simple by-products such as the creation of screenshots as well as screencast movies both annotated with the emulation's context information. The resulting IWD trace file has been reworked from previous versions making it both human readable (text-based) and efficient to parse and execute.

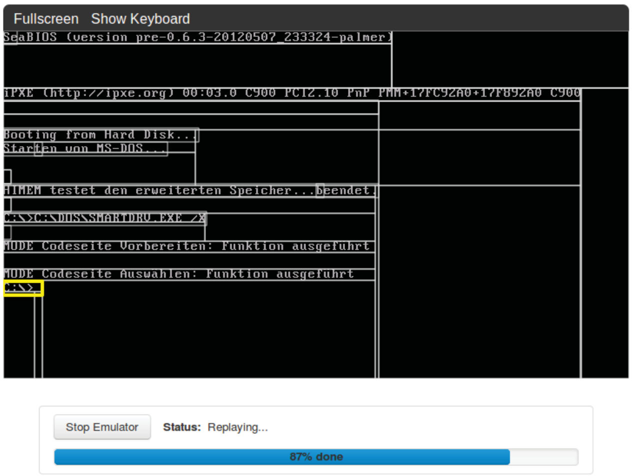
```
IWD := blocks{trace, meta-data, index}
trace :=  {<timestamp>|<instr_len>|<instr>}*
instr :=  <op_len>.<op>,<arg_length>.<arg>,...;
```

A *trace* block contains the session's events as an abstract description, such as the user's input-events, emulator output and synchronization data. For instance,
`520274|24|5.mouse,3.251,3.782,1.0;`
describes a mouse movement instruction (to 251,782) which occurred 520274*ns* after the start of the recording. The length of the instruction is 24 chars. For synchronization, events like:
`6226615|6434|5.vsync,2.13,2.77,2.40,2.30,6400.[...];`
describe pre- respectively postconditions. In this case 40×30 pixels at position $(13, 77)$ are expected to be similar to the Base64-encoded bitmap.
The trace file's *meta-data* section contains simple key/value pairs providing information regarding the trace file's creation context, e.g. a reference to the environment and emulator used as well as descriptive meta-data to be displayed. The trace file ends with an auto-generated *index* section that provides technical information for efficient parsing.

2.2 Replay

To replay an IWD, the trace-file is fed directly to the emulation component. If requested, the user is able to observe the emulator's visual output. For replay, also two worker threads are used, one for processing the trace file and another one for processing the emulator's output.
When replaying user interactions, it is possible that the emulator may drop events, e.g. if it cannot keep up with input processing. Furthermore, an action may take a varying amount of time for complete execution, such that recorded timestamps of events are not useable for input synchronization. Hence, the replay system has to adapt to the emulator's behavior. Since most input events produce a number of corresponding screen-updates[1], these updates, respectively the update patterns, are used for input synchronization and flow-control, i.e. delay processing of the next events until expected screen-regions are updated, hence the action's outcome is visible. Since screen-updates are not deterministic both by size or position, an expected update pattern is considered as successfully received if it covers the updated screen region. This way, visual synchronization is also available for environments without direct mouse input. Furthermore, this method is computationally efficient since no screen content has to be processed. Fig. 2 shows a recording of a console-based session. The yellow rectangle marks the screen area expected to be an outcome of an previous input action.

[1] For efficiency reasons only a set of tiles, covering changes on the screen are transferred.



```
SeaBIOS (version pre-0.6.3-20120507_233324-palmer)

iPXE (http://ipxe.org) 00:03.0 C900 PCI2.10 PnP PMM+17FC92A0+17F892A0 C900

Booting from Hard Disk...
Starten von MS-DOS...

HIMEM testet den erweiterten Speicher...beendet.

C:\>C:\DOS\SMARTDRV.EXE /X

MODE Codeseite Vorbereiten: Funktion ausgeführt

MODE Codeseite Auswählen: Funktion ausgeführt
C:\>_
```

Figure 2: Replay of a console based user interaction with visualized screen-updates (grey) and visual synchronization (yellow).

3. RESULTS, ISSUES & OUTLOOK

With a recording and replay system integrated into the EaaS framework, this system can be used with all available system environments. Yet, the system is usable for simple automation tasks as our tests still exposed unresolved issues, like failed or dropped mouse events (e.g. the mouse event had no effect and the window did not close). Furthermore, some animations, system-clock widgets caused problems due to non-deterministic screen-update events. These issues will be addressed by both implementing more robust pre- and post-condition checks as well as incorporating non-visual feedback from the emulation component (e.g. cpu and I/O status).
Other issues found in our tests, like unexpected error messages e.g. due to networking issues, missing menu-options or files, need to be addressed by the recording user.
Despite these yet unresolved issues, the presented system and its integration into EaaS workflows provides a stable base for new features for the digital preservation community to automate and document tasks on interactive systems. Furthermore, by unifying the communication with EaaS's emulation components, first steps for emulator-independent replay of interactions have been made. With this, captured interactions with today's emulators can be replayed with a future emulator hosting the same system environment.

4. REFERENCES

[1] K. Rechert, I. Valizada, D. von Suchodoletz, and J. Latocha. bwFLA – A Functional Approach to Digital Preservation. *PIK – Praxis der Informationsverarbeitung und Kommunikation*, 35(4):259–267, 2012.

[2] K. Rechert, D. von Suchodoletz, R. Welte, M. van den Dobbelsteen, B. Roberts, J. van der Hoeven, and J. Schroder. Novel workflows for abstract handling of complex interaction processes in digital preservation. In *Proceedings of the 6th International Conference on Preservation of Digital Objects (iPRES2009)*, pages 155–161, 2009.

Reviving Antique Software: Curation Challenges and the Olive Archive

Daniel Ryan
Carnegie Mellon University
WEH 4418, 5000 Forbes Ave
Pittsburgh PA 15213
+1 (412) 268-5278
dfryan@andrew.cmu.edu

Gloriana St. Clair
Carnegie Mellon University
WEH 4418, 5000 Forbes Ave
Pittsburgh PA 15213
+1 (412) 268-5278
gstclair@andrew.cmu.edu

ABSTRACT

A growing percentage of the world's intellectual output is in the form of executable content, such as simulation models, tutoring systems, data visualization tools, and expert systems. To preserve this content over time, we need to freeze and precisely reproduce the execution that dynamically produces that content. Olive, a rough acronym for "Open Library of Images for Virtualized Execution," is a system built at Carnegie Mellon University. Olive preserves and provides access to this executable content. It relies on virtual machine (VM) technology to bundle software with all of its dependencies. These VMs are streamed over the internet in real time to ensure a smooth user experience while maintaining fidelity to the original execution environment[1].

This demonstration examines some of the challenges the Olive team has encountered in the process of preserving software over the last several years. Among these difficulties are technical challenges, problems of scale, legal limitations, and a lack of existing curation standards for executable content.

General Terms

infrastructure, preservation strategies and workflows, specialist content types, case studies and best practice.

Keywords

preservation, software, virtualization.

1. INTRODUCTION

Born-digital interactive content makes up an increasing proportion of creative and scholarly output around the world today. The global, instantaneous, and unrestrainable nature of software has made it a major part of our cultural heritage. Significantly, executable content draws its cultural impact from its interactivity: users have to participate and interact with software in order to understand what it does, how it works, and why it is useful.

Historically, libraries, museums, and other cultural memory organizations have been effective in preserving the developing record of civilization globally, and in assisting the users of that record to understand it and to use it to create new knowledge. In the arts and humanities, citizens and scholars can view cave paintings at Lascaux, the Bayeux tapestry, the Bill of Rights, the archival papers of U.S. Senator John Heinz, and over twenty million books. Published scholarly work is more widely disseminated than it has ever been. Those interested in their heritage can listen to traditional music, study ancient commercial records and texts, and attend plays written by Shakespeare. Currently, these seekers cannot use primary source materials from the growing realm of executable content, because the *execution environment* is not compatible with modern technology. Instead, scholars must rely on a variety of secondary sources, including screenshots, descriptions, and community commentary.

In *Preserving Digital Information, Report of the Task Force on Archiving of Digital Information,* Don Waters and John Garrett made a daunting prediction that if libraries did not seek to preserve digital information, the result would be difficult. "Failure to look for trusted means and methods of digital preservation will certainly exact a stiff, long-term cultural penalty[2]."

The pervasiveness of executable content is a worldwide phenomenon. When historians look back on the nature of society during the computer revolution, they will need working, perfectly faithful instances of the software in use and the experience of interacting with it. When sociologists seek to understand exactly which characteristics of Angry Birds drove many adults internationally to spend large portions of time flinging digital birds at digital houses, they will need to run it and play it themselves. No explanation or description could suffice.

2. PROJECT HISTORY

In 2012, Carnegie Mellon computer science professor Mahadev Satyanaranan (Satya) approached the Dean of Libraries, Gloriana St. Clair, to discuss a project for which he saw an application that might be suited to the University Libraries. Satya had been working with Vasanth Bala (Vas) at IBM Research to package and stream virtual machines for fast application deployment.. This project was known as Internet Suspend/Resume® (ISR). As the project evolved, Satya and Vas began to see ISR's potential for preserving software. The ISR team understood the technical and infrastructural challenges behind such a project, and thought it was worth investing the time and money to devise a solution. Neither IBM nor Satya was interested, however, in keeping old things around forever. They agreed to begin by reaching out to the Carnegie Mellon Libraries, where Gloriana had established a reputation as a digital pioneer and an extensive collaborator with the computer science department. Thus, the Olive project was born. St. Clair assured Satya that not only were she and the CMU Libraries interested in solving this problem, but also that the library community shared her sense of responsibility around executable content.

In 2010, IBM hosted a meeting to test the idea that libraries and campus computing might be interested in preserving executable content. Participants were enthusiastic about the technology, anxious about the legal situation, and worried about both the economic and the organizational issues.

Both IMLS and the Sloan Foundation gave grants for a proof of concept phase of Olive development. Since October 2012, the Olive project has received $497,000 from the Institute for Museum and Library Services, and $400,000 from the Sloan Foundation, to support a proof of concept effort and development. Part of the funding sought from the Sloan Foundation was awarded to Ithaka S+R for a whitepaper on sustaining an entity like Olive after the core research and development has been completed. The report recommends an additional three years of funding for intensive R&D, followed by the formation of a sustaining coalition of interested parties sharing the financial burden of the operational costs of such an archive.

3. APPROACH

3.1 Execution Fidelity

Software reproduction is a complex problem, the solution of which requires the perfect alignment of many moving parts. Achieving execution fidelity has long evaded preservationists and has stymied the efforts of the digital library community to archive executable content[8][3][4]. Even minor changes can cause a breakdown in the stability of the execution environment. These changes can include dynamically linked libraries, preferences, configuration files, clock timings, hardware capabilities, and more. Simply constructing the appropriate environment in which legacy software will execute often requires expert knowledge of the original environment. We refer to the successful alignment of all of these variables as *execution fidelity*[5][6]. As legacy software falls further into deprecation, the level of knowledge required to achieve execution fidelity becomes increasingly rare.

3.2 Virtual Machines

In order to encapsulate an execution environment, Olive relies on virtual machine technology. Communicating with a *virtual machine monitor*, VM images are supplied with a virtualized representation of a computer architecture and instruction set see *Figure 1*). Virtual machine monitors leverage the actual hardware of a machine (the host machine) to ensure that the operating system and applications inside are unable to distinguish between the virtual environment and a real legacy system. This precise imitation of hardware is why Olive relies on virtualization as a preservation strategy.

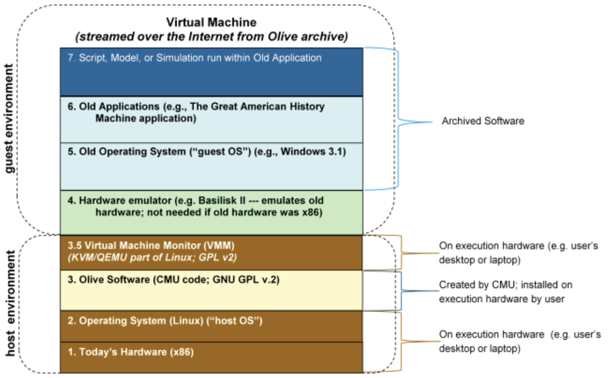
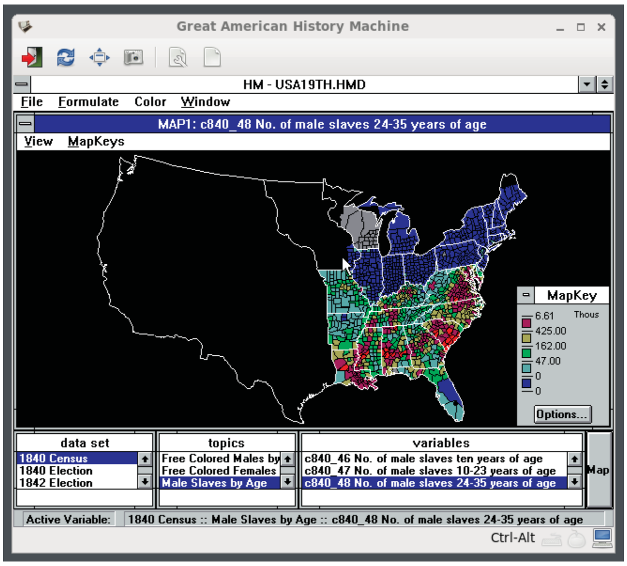

Virtual Machine
(streamed over the Internet from Olive archive)
guest environment
7. Script, Model, or Simulation run within Old Application
6. Old Applications (e.g., The Great American History Machine application)
5. Old Operating System ("guest OS") (e.g., Windows 3.1)
4. Hardware emulator (e.g. Basilisk II --- emulates old hardware; not needed if old hardware was x86)
Archived Software
host environment
3.5 Virtual Machine Monitor (VMM) (KVM/QEMU part of Linux; GPL v2)
3. Olive Software (CMU code; GNU GPL v.2)
2. Operating System (Linux) ("host OS")
1. Today's Hardware (x86)
On execution hardware (e.g. user's desktop or laptop)
Created by CMU; installed on execution hardware by user
On execution hardware (e.g. user's desktop or laptop)

Figure 1: Olive Client Architecture



Active Variable: 1840 Census :: Male Slaves by Age :: c840_48 No. of male slaves 24-35 years of age

Figure 2: Great American History Machine (Windows 3.1)

Olive is built on standard, unmodified web technologies (standard web servers, HTTP for communication) and works to stream VM images in pieces as they are requested. Execution can happen either directly on a user's computer or on a compute node dedicated to VM execution.

3.3 Examples & Demo

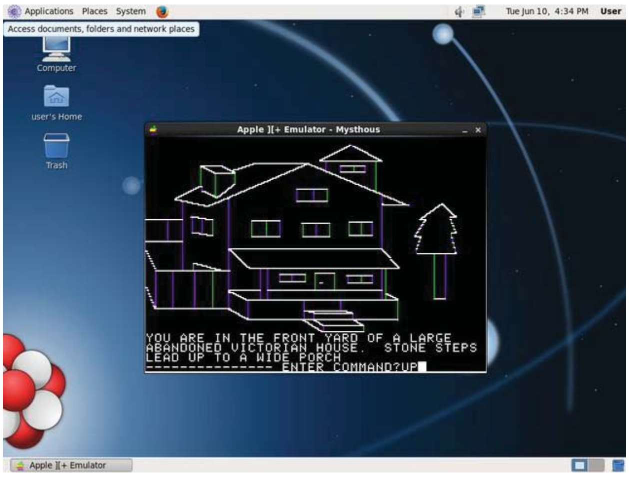
There are several pieces of software archived in Olive, but here we will focus on only two brief examples:

1. The Great American History Machine (see Figure 2): A piece of educational software written in the late 1980's by Professor David Miller at Carnegie Mellon. This software was used to teach early American History at institutions across the United States. It offers unique tools for exploring census and election data. Professor Miller and his team did not have the technical resources to migrate this tool when Windows 3.1 became deprecated, so the software fell into disuse until we recovered it.
2. Mystery House (see Figure 3): Mystery House is the original graphically-rendered adventure game written for the Apple II. It brought graphical interaction to the mainstream just over 30 years ago, yet actually running that software today is a significant challenge; not only did we need the original disk image, but we also had to find an Apple II emulator and the accompanying ROM (read-only memory), which was originally built into the machine. Without archival, executable instances of software like Mystery House, we lose our ability to reflect on the history of computer games and human/computer interaction.

These examples highlight the potential for olive to preserve and provide access to software which might otherwise be lost.

4. CHALLENGES

4.1 Technical Challenges



Applications Places System
Tue Jun 10, 4:34 PM User
Access documents, folders and network places
Computer
user's Home
Trash
Apple][+ Emulator - Mysthous
Apple][+ Emulator

```
YOU ARE IN THE FRONT YARD OF A LARGE
ABANDONED VICTORIAN HOUSE.  STONE STEPS
LEAD UP TO A WIDE PORCH
------------------ ENTER COMMAND?UP
```

Figure 1: Mystery House on Apple II emulator

In the simplest terms, Olive will be like YouTube for executable content. Olive provides a tool for preserving and remotely accessing software. To preserve this content over time, we need to freeze and precisely reproduce the execution that dynamically produces the content. While this may sound simple, many have studied the problem over the last two decades, but only now are successful efforts underway.

Here are a few of the technical challenges we have encountered while trying to achieve a working implementation of Olive:

- Low latency streaming and caching of VM images[8];
- Lack of backward compatibility in updated releases of dependent software;
- Bugs which existed in old software/hardware but only present themselves in modern systems;
- Effective, secure, and flexible implementation of access controls, and
- Failure of modern VMMs to represent faithfully the extended memory space required to run certain systems.

For example, the version of qemu/kvm bundled with Ubuntu 12.04 was several releases out of date as compared with that packaged with Redhat Enterprise Linux. VMs built on RHEL 6.x would fail to boot when exported to an Ubuntu 12.04 machine with qemu/kvm installed from the normal Ubuntu repository. In order to overcome this difficulty, the software Olive provides for packaging VMs strips down and validates the XML. This XML is responsible for defining the configuration of a VM in order to ensure continuing compatibility, both forward and backward[7][8].

In another edge case, we discovered that Windows 3.1 mouse support suffered from a strange bug which caused the mouse pointer to jump randomly around the screen. Upon investigation, the Olive development team learned that the serial mouse drivers for Windows 3.1 contain an off-by-one error which is only exposed when mouse updates occur more than 40 times per second. On older mice, this did not cause problems because they did not send updates so frequently. However, modern laser mice send information much more often than 40 times per second. After tracking this issue down, we were forced to construct a binary patch for the driver.

4.2 Legal Challenges

The world's accrued wisdom is available to scholars and students globally. In general, the pre-1923 U.S. content can be benefited from without much concern about being sued for reusing that work in the creation of newer work. For instance, Shakespeare's output can be performed in all kinds of redacted and enhanced formats and interpretations. Shakespeare's heirs may wince but they cannot and do not sue. In striking contrast, the family-profit-maximizing Tolkien Estate manages its assets by aggressively controlling all aspects in all formats. J. R. R. Tolkien himself sued Ace Books for publishing a pirated paperback edition of *The Lord of the Rings*. Ace paid damages and Tolkien's publisher moved to meet demand by bringing out an authorized paperback. The Tolkien Estate continues to be zealous in managing its property. For a less popular author, this approach might be detrimental.

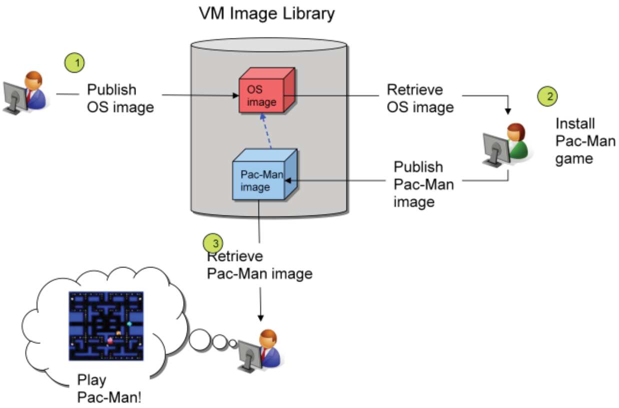


Figure 2: Crowdsourced Publication Workflow

Generally, most post-1923 content requires some kind of license in most countries. Presently, Olive is a closed research project, which affords it certain protections from infringement claims under fair use provisions of the copyright law. However, we recognize the need for an open, accessible archive for software, and CMU General Counsel Mary Jo Dively commissioned an extensive risk assessment of varying levels of public access to Olive. We are continuing to study this report.

4.3 Curation Challenges

Many collections of historical material are established, managed, and maintained by curators, who are responsible for selecting content, developing and applying an acquisition process, and keeping that content secure from degradation. Often this means protecting works of art from sunlight and flash photography, or protecting books from falling apart. When the object of curation is a piece of software in executable form, however, the process of curation is not particularly well defined. For a given piece of software, curation might involve identifying the hardware it requires to operate, locating an emulator for that hardware (if necessary), determining the platform and version required to run the software, configuring the emulator, installing and configuring the platform, locating and importing dependent drivers, installing the software, ensuring faithful behavior, generating metadata, and tracking down related rights information, and packaging and uploading the containing VM.

This set of tasks would be daunting enough given a modern, well documented technology stack. For old or deprecated software, the dependency stack will often require extensive expertise to configure and install successfully, if it is still possible to identify and acquire the full dependency stack at all. Documentation for these configuration and installation procedures is often lacking, and finding an expert will become increasingly difficult.

As part of a grant from the Institute for Museum and Library Services, the Olive team agreed to preserve Doom, the original first-person shooter game written for MS-DOS. Beginning with an image of the original MS-DOS 6.22 installation floppy disk, we soon learned that reliable instructions for achieving a successful system configuration were scarce, poorly documented, and largely dependent upon third party additions with similar challenges. Similar issues arose when we attempted to install Windows 95.

Because of the degree of expertise required and the sheer quantity of software which is in jeopardy of becoming extinct, we currently plan to investigate crowd-sourced curation models in the next phase of our work. As we move forward, our development team is implementing functionality to allow new VMs to be published as a changeset applied to an existing VM, which would eliminate the need to confront a complex dependency stack more than once. A sample curation workflow supported by this model can be seen in Figure 4.

5. CONCLUSION

Preserving software in its execution environment is critically important to our institutional goal of preserving the cultural record. The Olive Archive is an infrastructure designed to limit challenges to future curators, but will begin to rely more heavily on community involvement in the coming years. Many important questions must be addressed by curators and preservation experts as institutions take on the daunting challenge of capturing, describing, checking, cleaning, migrating, and maintaining collections of software in virtual machines.

6. ACKNOWLEDGMENTS

The Olive Archive is supported by grant funding from the Institute for Museum and Library Services and from the Sloan Foundation. We are grateful to Vas Bala and IBM for initially supporting this research, to Carnegie Mellon for financial, legal, technical and moral support, and to Deanna Marcum and Ithaka S+R for their ongoing insight and assistance. Final thanks to the advisory group and to our project team: Erika Linke, Mahadev Satyanarayanan, Keith Webster, Benjamin Gilbert, Jan Harkes, Jerome McDonough, and Anita de Waard.

7. REFERENCES

[1] Open source software at the Olive Archive can be found available at https://github.com/cmusatyalab.

[2] Donald Waters and John Garrett, "Preserving Digital Information, Report of the Task Force on Archiving of Digital Information," Council on Library and Information Resources, May 1996. Available: http://www.clir.org/pubs/abstract//reports/pub63.

[3] P. Conway. Preservation in the Digital World. http://www.clir.org/pubs/reports/conway2/, March 1996..

[4] P. Conway. Preservation in the Age of Google: Digitization, Digital Preservation, and Dilemmas. *Library Quarterly*, 80(1), 2010.

[5] B. Matthews, A. Shaon, J. Bicarreguil, and C. Jones. A Framework for Software Preservation. *The International Journal of Digital Curation*, 5(1), June 2010.

[6] Satyanarayanan, Mahadev ; Bala, Vasanth ; Clair, Gloriana St. ; Linke, Erika ; Georgakopoulos, Dimitrios (Bearb.) ; Joshi, James B. D. (Bearb.): Collaborating with executable content across space and time.. In:*CollaborateCom* : IEEE, 2011. - ISBN 978-1-4673-0683-6, S. 528-537.

[7] Gilbert, Benjamin. 2013. Building VMNetX with qemu and libvirt. Workshop. Carnegie Mellon University (Jun. 2013), https://olivearchive.org/static/documents/vmnetx-gilbert.pdf.

[8] Bowman, M., Debray, S. K., and Peterson, L. L. 1993. Reasoning about naming systems. *ACM Trans. Program. Lang. Syst.* 15, 5 (Nov. 1993), 795-825. DOI= http://doi.acm.org/10.1145/161468.16147.

[9] Yoshihisa Abe, Roxana Geambasu, Kaustubh Joshi, H. Andrés Lagar-Cavilla, and Mahadev Satyanarayanan. 2013. vTube: efficient streaming of virtual appliances over last-mile networks. In *Proceedings of the 4th annual Symposium on Cloud Computing* (SOCC '13). ACM, New York, NY, USA, , Article 16 , 16 pages. DOI=10.1145/2523616.2523636 http://doi.acm.org/10.1145/2523616.2523636

NLA Software and File Formats Knowledge Base

Dr Mark Pearson
National Library of Australia
Parkes Place West
Canberra, ACT 2600
Australia
+61 2 6262 1080
mark.pearson@nla.gov.au

Gareth Kay
National Library of Australia
Parkes Place West
Canberra, ACT 2600
Australia
+61 2 6262 1031
gareth.kay@nla.gov.au

ABSTRACT

This demonstration will showcase ongoing work at the National Library of Australia to develop a software and file formats knowledge base for digital preservation purposes. This project involves empirical research into the capabilities of software applications in relation to file formats. We will talk about the types of information we capture in the knowledge base and describe the steps we are taking to transform it into a machine-actionable graph database, a prototype of which will also be demonstrated.

General Terms

infrastructure

Keywords

Software, file format, knowledge base, graph database.

1. INTRODUCTION

The National Library of Australia has an ongoing project to develop a knowledge base detailing relationships between software applications and file formats. This paper describes the drivers and strategic goals for developing the knowledge base and the rationale behind taking an empirical approach to its development.

2. SCOPE OF THE WORK

The project involves detailed empirical research into the capabilities of selected software applications with respect to selected file formats. The research is primarily format-driven since the primary long-term goal is to be able to successfully access content stored in digital files. Priority file formats have been chosen based on business needs and the composition of the NLA's digital collections.

For each major abstract content type (images, textual documents, videos, spreadsheets, maps etc.) we have chosen a small number of the most predominantly used applications in order to investigate their capabilities with respect to the associated file formats. The applications chosen may be proprietary in nature or open source.

Details such as release dates, versions, vendor support, licensing status and dependencies are recorded both for formats and applications. Due to business needs the data gathered from the research is initially being recorded in a multiple worksheet Excel file in semi-structured format. Development of a prototype graph database together with software modules capable of importing data from the Excel file is taking place in parallel with the empirical work.

While Excel is not a suitable platform for a production knowledge base, its use in the development phase does have some advantages: as our understanding of the problem domain improves through empirical contact with it, we can experiment with changes to our data model at very little cost. When we come across aspects of the software/file-format relationship which we judge might be significant to future preservation decision making but which the current iteration of the model provides no structured way to record, we can adapt the model accordingly.

Two very useful by-products of the empirical work are: a growing corpus of files in various formats and format versions containing known content which we have created ourselves and which we can usefully employ in testing software package capabilities; and a growing collection of VMWare virtual machine images for various current and historical operating system environments.

3. GOALS AND DRIVERS

The long-term strategic goal is to build machine-readable knowledge bases to aid us in: determining our *level of support* [1] for different file formats; analysing the NLA's digital collection materials for preservation risks; and planning and executing preservation actions on digital objects which comply with the documented *preservation intents* [2] for those objects.

A more immediate goal is to replace an existing Drupal-based software/formats knowledge base which is limited in its ability to express arbitrary relationships between entities and is not suitable for machine querying or complex queries.

There is much existing work in the area of technical registries [3][4][5][6][7] and the NLA is actively involved in other work in this area through collaboration with organisations such as National and State Libraries Australasia [8]. While the outcomes of this project will provide practical benefits for the NLA they will also hopefully provide food for thought for the wider community in its efforts to develop open, maintainable linked data technical registries.

4. THE KNOWLEDGE BASE

Functional relationships – A key function of the knowledge base is to map out the capabilities of software applications in relation to the file formats they are (or claim to be) able to handle. To gather this data we investigate certain *functional relationships* for each software/format combination. These relationships are used to describe capabilities exhibited by an application in relation to a format. Currently, we investigate four relationships: *import*; *render*; *edit*; and *save*. These relate to whether an application can parse a given format and build a 'meaningful' internal representation of its content; render that internal representation; allow a user to make changes to the content; and save it to the format, respectively.

The process of documenting these functional relationships involves as a first step harvesting information (where available) from vendor documentation and/or running the software and noting the formats listed in the 'Open' and 'Save as' menus. Such entries in the knowledge base are assigned a confidence value of 'untested' as we don't know how *well* the software opens or saves a given format. For file formats which are considered high priority by the NLA the functional relationships are empirically tested with the aid of the *test file corpus* (described below). Such entries are assigned a confidence value of 'tested' and are more detailed in nature.

Preservation notes – During the process of investigating the functional relationships issues which could affect the suitability of either an application or a file format for future preservation actions sometimes arise. Examples could include rendering issues; discrepancies between documented and actual software functionality; software/hardware dependencies; installation issues; and/or the inability to preserve certain properties of content which may have been deemed significant by the preservation intent statements associated with the content type. These issues are recorded in semi-structured format in the 'preservation notes' field. Crucially these notes can act as triggers for reassessing the knowledge base schema if it becomes clear that the current schema provides no structured means for recording such details.

Test file corpus – A useful by-product of this project is a growing benchmark corpus of test files in selected file formats, created in software packages which have been documented in the knowledge base. When new test files are created, their content is carefully crafted in accordance with current preservation intent statements for the content type. Put more simply, the content is chosen so that we can test how well an application maintains important features of that content when importing, rendering, editing or saving it.

What makes this corpus particularly valuable is that each test file is linked within the knowledge base to the software version used to create the test file, the operating system and environment in which the software was run, as well as the format and version the file is written in. Another feature is the inclusion of screenshots showing the content from each file rendered in the software it was created with. This additional resource allows for the detection of content loss or rendering issues when a file is opened in a different application.

5. GRAPH DATABASE

When the empirical part of this work began we did not have a suitable database in which to record our findings and for this reason, as mentioned above, the data is currently recorded in an Excel file in semi-structured format. However, in parallel with the empirical work we have also been developing a set of software modules for importing and transforming the Excel data into a *directed property graph*: a "key/value-based, directed, multi-relational graph" [9]. In such graphs both vertices and edges may have arbitrary sets of key/value attributes.

A number of database systems supporting the property graph model are currently available [10] but the system we have chosen initially – OrientDB [11] – supports vertex and edge types which can have inheritance relationships. Vertices, edges and vertex/edge types can all be dynamically added and removed. This makes it ideally suited to problem domains where the schema has not been strictly defined or may be a 'moving target'. It is also open source, released under the Apache License, Version 2.0.

OrientDB supports the Tinkerpop Blueprints property graph Java API [10] - described as "JDBC, but for graph databases" and both an SQL-based query language extended with features for graph traversal; and Gremlin [11] – a graph traversal language.

6. THE DEMONSTRATION

The demonstration at iPres 2014 will address in more detail the data we are capturing in the knowledge base and the nature of the functional relationships we test for each software/format combination. The prototype graph database will also be demonstrated with example queries.

7. REFERENCES

[1] Pearson, D. 2012. The Adventures of Digi: Ideas, Requirements and Reality. Presentation at *Future Perfect 2012*, Museum of New Zealand Te Papa Tongarewa, Wellington. https://www.nla.gov.au/content/the-adventures-of-digi-ideas-requirements-and-reality

[2] Webb, C., Pearson, D. and Koerbin, P. 2013. *"Oh, you wanted us to preserve that?!"* Statements of Preservation Intent for the National Library of Australia's Digital Collections'. D-Lib Magazine, Vol.19 1/2, Jan/Feb 2013.

[3] *PRONOM technical registry* - http://apps.nationalarchives.gov.uk/pronom/

[4] *Unified Digital Format Registry (UDFR)* - http://www.udfr.org/

[5] Anderson, David and Delve, Janet (2012) *The Trusted Online Technical Environment Metadata Database –* TOTEM. In: Kölner Beiträge zu einer geisteswissenschaftlichen Fachinformatik . Verlag Dr. Kovac, Hamburg. ISBN 9783830064183

[6] http://www.openplanetsfoundation.org/blogs/2010-12-08-breaking-down-format-registry

[7] McGath, G. 2013 *The Format Registry Problem.* Code4Lib Journal. Issue 19, 2013-01-15. http://journal.code4lib.org/articles/8029

[8] http://www.nsla.org.au/projects/digital-preservation

[9] *Defining a Property Graph* - https://github.com/tinkerpop/gremlin/wiki/Defining-a-Property-Graph

[10] *Tinkerpop Blueprints* home page - https://github.com/tinkerpop/blueprints/wiki

[11] *Orient Technologies* home page - http://www.orientechnologies.com

[12] *A Graph Traversal Language* - https://github.com/tinkerpop/gremlin/wiki