

BRITISH
LIBRARY

iPRES 2008

Proceedings of The Fifth International
Conference on Preservation of Digital Objects
Joined Up and Working: Tools and Methods
for Digital Preservation

The British Library, London. 29 – 30 September



**The British Library
St Pancras
London, UK**

©The British Library Board. The rights of the authors contained in this publication to be identified as the author of the work have been asserted by him or her in accordance with the Copyright & Designs Patent Act 1988. All rights reserved.

Foreword

This volume brings together the proceedings of iPRES 2008, the Fifth International Conference on Digital Preservation, held at The British Library on 29-30 September, 2008. From its beginnings five years ago, iPRES has retained its strong international flavour. This year, it brings together over 250 participants from 33 countries and four continents. iPRES has become a major international forum for the exchange of ideas and practice in Digital Preservation.

The theme of the conference is 'Joined Up and Working: Tools and Methods for Digital Preservation'. Preserving our scientific, cultural and social digital heritage draws together activity across diverse disciplines. It transcends international boundaries and incorporates the needs of disparate communities. By working together, we have been able to make real concrete progress towards solving the problems that we identified in earlier years.

The opening address by Dame Lynne Brindley, CEO of The British Library, demonstrates both the importance of Digital Preservation at the national level, as well as commitment to dedicate the resources needed to make progress.

The iPRES 2008 conference theme and the papers gathered together here represent a major shift in the state-of-the-art. For the first time, this progress enabled the Programme Committee to establish two distinct tracks. The practitioner track is designed for those with an interest in practically preserving digital content within their organisation. The technical track is designed for those with an interest in underpinning concepts and digital preservation technology. Readers will find valuable insights to draw from in both areas.

This is also the first year that iPRES has collected and published full papers in addition to the presentations provided at the conference. Authors' abstracts were reviewed by at least three members of the Programme Committee for quality, innovation, and significance. The Programme Committee was impressed by the high quality of the submissions. The best 50 were invited to provide full papers for inclusion in the proceedings and presentation at the conference. There are a very limited number of venues for publishing conceptual frameworks, scientific results, and practical experience in Digital Preservation. I believe that inclusion of full papers will make an important contribution to the field by addressing this problem.

It is a huge effort to organise a successful international conference. Thanks are due to many individuals and organisations. In particular, we thank the members of the Programme Committee, the members of the Organising Committee, the invited speakers, panellists, authors, presenters, and the participants. We are also grateful for the support provided by The British Library, the Digital Preservation Coalition (DPC) and JISC. In addition, we thank Ex Libris, Sun Microsystems and Tessella for their recognition of the challenge presented by the long term preservation of digital content and their support for this conference.

Dr. Adam Farqhar

Programme Chair

Table of Contents

iPRES2008 Programme and Organising Committee

Paper Session 1: Modelling Organisational Goals

Session Chair: Oya Rieger (*Cornell University*)

- **Digital Preservation Policy: A Subject of No Importance?** 1
Neil Beagrie, Najla Rettberg, Peter Williams (*Charles Beagrie Ltd*)
- **Modelling Organisational Preservation Goals to Guide Digital Preservation** 5
Angela Dappert, Adam Farquhar (*The British Library*)
- **Component Business Model for Digital Preservation: A Framework for Analysis** 13
Raymond Van Diessen (*IBM*), Barbara Sierman (*National Library of the Netherlands*), Christopher Lee (*University of North Carolina*)
- **Development of Organisational and Business Models for the Long-term Preservation of Digital Objects** 20
Tobias Beinert, Suzanne Lang, Astrid Shoger (*Bavarian State Library*)
Uwe Borghoff, Harald Hagel, Michael Minkus, Peter Rödiger (*University of Federal Armed Force*)

Paper Session 2: Disciplinary Contexts

Session Chair: John Kunze (*California Digital Library, University of California*)

- **Long-term Preservation of Electronic Literature** 28
Sabine Schrimpf (*German National Library*)
- **Preservation of Art in the Digital Realm** 32
Tim au Yeung, Sheelagh Carpendale, Saul Greenberg (*University of Calgary*)
- **In Cypher Writ, or New Made Idioms: Sustaining Digital Scholarship as Cooperative Digital Preservation** 40
Bradley Daigle (*University of Virginia*)
- **Adapting Existing Technologies for Digitally Archiving Personal Lives: Digital Forensics, Ancestral Computing and Evolutionary Perspectives and Tools** 48
Jeremy Leighton John (*The British Library*)

Paper Session 3: Digital Preservation Formats

Session Chair: Adrian Brown (*The National Archives*)

- **Enduring Access to Digitised Books: Organizational and Technical Framework** 56
Oya Rieger, Bill Kehoe (*Cornell University*)
- **Creating Virtual CD-Rom Collections** 62
Kam Woods, Geoffrey Brown (*Indiana University*)
- **Preservation of Web Resources: The JISC-PoWR Project** 70
Brian Kelly, Marieke Guy (*UKOLN, University of Bath*)
Kevin Ashley, Ed Pinsent, Richard Davis (*University of London Computer Centre*)
Jordan Hatcher (*Opencontentlawyer.com*)

- **Preserving the Content and Network: An Innovative Approach to Web Archiving** 78
Amanda Spencer (*The National Archives*)
- **“What? So What?”: the Next-Generation JHOVE2 Architecture for Format-Aware Characterization** 86
Stephen Abrams (*California Digital Library, University of California*)
Sheelagh Morrissey (*Portico*)
Tom Cramer (*Stanford University*)

Paper Session 4: Preservation Planning

Session Chair: Kevin Ashley (*University of London Computer Centre*)

- **Emulation: From Digital Artefact to Remotely Rendered Environments** 93
Dirk von Suchodoletz, Jeffrey van der Hoeven (*University of Freiburg*)
- **Data Without Meaning: Establishing the Significant Properties of Digital Research** 99
Gareth Knight (*Kings College, London*), Maureen Pennock (*Portico*)
- **Towards a Curation and Preservation Architecture for CAD Engineering Models** 107
Alexander Ball, Manjula Patel, Lian Ding (*University of Bath*)
- **Evaluating Strategies for Preservation of Console Video Games** 115
Mark Guttenbrunner, Christoph Becker, Andreas Rauber, Carmen Kehrborg (*Vienna University of Technology*)

Paper Session 5: Understanding Costs & Risks

Session Chair: Neil Beagrie (*Charles Beagrie Ltd*)

- **Costing the Digital Preservation Lifecycle More Effectively** 122
Paul Wheatley (*The British Library*)
- **Risk Assessment: Using a Risk-based Approach to Prioritise Handheld Digital Information** 127
Rory McLeod (*The British Library*)
- **The Significance of Storage in the ‘Cost of Risk’ of Digital Preservation** 134
Richard Wright, Art Miller (*BBC*), Matthew Addis (*University of Southampton*)
- **International Study on the Copyright and Digital Preservation** 140
June Besek (*Columbia Law School*)
Jessica Coates, Brian Fitzgerald (*Queensland University of Technology*)
William LeFurgy, Christopher Weston (*Library of Congress*)
Wilma Moussink (*SURFfoundation*)
Adrienne Muir (*University of Loughborough*)

Paper Session 6: Preservation Metadata

Session Chair: Michael Day (*UKOLN, University of Bath*)

- **Developing Preservation Metadata for Use in Grid-based Preservation Systems** 145
Arwen Hutt, Brad Westbrook, Ardys Kozbial (*University of California*)
Robert McDonald (*Indiana University*)
Don Sutton (*San Diego Super Computer Center*)

- **Using METS, PREMIS and MODS for Archiving EJournals** 151
Angela Dappert, Markus Enders (*The British Library*)
- **Harvester Results in Digital Preservation System** 159
Tobias Steinke (*German National Library*)
- **The FRBR-Theoretical Library: The Role of Conceptual Data Modelling in Cultural Heritage Information System Design** 163
Ronald Murray (*Library of Congress*)

Paper Session 7: Grid Storage Architecture

Session Chair: Steve Abrams (*California Digital Library, University of California*)

- **Towards Smart Storage for Repository Preservation Services** 169
Steve Hitchcock, David Tarrant, Leslie Carr (*University of Southampton*)
Adrian Brown (*The National Archives*)
Ben O'Steen, Neil Jefferies (*Oxford University*)
- **Repository and Preservation Storage Architecture** 175
Keith Rajewski (*Sun Microsystems Inc.*)
- **Implementing Preservation Services over the Storage Resource Broker** 181
Douglas Kosovic, Jane Hunter (*University of Queensland*)
- **Embedding Legacy Environments into a Grid-based Preservation Infrastructure** 189
Claus-Peter Klas, Holger Brocks, Lars Müller, Matthias Hemmje (*Fern Universität, Hagen*)

Paper Session 8: Establishing Trust in Service Providers

Session Chair: Seamus Ross (*Humanities Advanced Technology and Information Institute*)

- **Creating Trust Relationships for Distributed Digital Preservation Federations** 197
Tyler Walters (*Georgia Institute of Technology Library and Information Centre*)
Robert McDonald (*Indiana University*)
- **The Use of Quality Management Standards in Trustworthy Digital Archives** 205
Susanne Dobratz, (*Humboldt University, Berlin*)
Uwe Borgoff, Peter Rödiger (*University of the Federal Armed Forces*)
Astrid Schoger (*Bavarian State Library*)
Björn Rätzke – (*Rätzke IT Services*)
- **The Data Audit Framework: A Toolkit to Identify Research Assets and Improve Data Management in Research-led Institutions** 213
Sarah Jones, Seamus Ross (*Digital Curation Centre and Humanities Advanced Technology and Information Institute*)
Raivo Ruusalepp (*Estonian Business Archives*)
- **Data Seal of Approval – Assessment and Review of the Quality of Operations for Research Data Repositories** 220
Henk Harmsen (*Data Archiving and Networked Services*)

Paper Session 9: Service Architectures for Digital Preservation

Session Chair: Andrew Wilson (*National Archives of Australia*)

- **Updating DAITSS - Transitioning to a Web Service Architecture** 223
Randall Fischer, Carol Chou, Franco Lazzarino (*Florida Center for Library Automation*)
- **Conceptual Framework for the Use of the Service-oriented Architecture-Approach in the Digital Preservation** 229
Christian Saul, Fanny Klett (*Fraunhofer Institute of Digital Media Technology*)
- **RODA and Crib: A Service Oriented Digital Repository** 235
José Carlos Ramalho, Miguel Ferreira (*University of Minho*)
Luis Faria, Rui Castro, Francisco Barbedo, Luis Corugo, ((*DGARQ*))
- **Persistent Identifier Distributed System for Cultural Heritage Digital Objects** 242
Emanuele Bellini, Chiara Cirinnà, Maurizio Lunghi (*Foundation Rinascimento Digitale*)
Ernesto Damiani, Cristiano Fugazza (*University of Milan*)

Paper Session 10: Digital Preservation Services

Session Chair: John Kunze (*California Digital Library, University of California*)

- **Encouraging Cyberinfrastructure Collaboration for Digital Preservation** 250
Christopher Jordan (*Texas Advanced Computing Center*)
Ardys Kozbial (*University of California*)
David Minor, Robert McDonald (*San Diego Super Computer Centre*)
- **Establishing a Community-based Approach to Electronic Journal Archiving: LOCKSS** 257
Adam Rusbridge, Seamus Ross (*Digital Curation Centre*)
- **The KB e-Depot in Development Integrating Research Results in the Library Organisation** 264
Hilde van Wijngaarden, Frank Houtman, Marcel Ras (*National Library of the Netherlands*)
- **Building a Digital Repository: a Practical Implementation** 270
Filip Boudrez (*City Archives of Antwerp*)

Paper Session 11: Foundations

Session Chair: Chris Rusbridge (*Digital Curation Centre*)

- **Bit Preservation: A Solved Problem?** 274
David H Rosenthal (*Stanford University*)
- **The Modelling System Reliability for Digital Preservation: Model Modification and Four-Copy Model Study** 281
Yan Han, Chi Pak Chan (*University of Arizona*)
- **Ways to Deal with Complexity** 287
Christian Keitel (*Staatsarchiv Ludwigsburg*)

292

- **A Logic-based Approach to the Formal Specification of Data Formats**
Michael Hartle, Arsenne Botchak, Daniel Schumann, Max Mühlhäuser (*Technische Universität, Darmstadt*)

Panels:

- **National and International Initiatives Panel Discussion** 300
Moderator: Neil Grindley (*Joint Information Systems Committee, JISC*)
Panelists: Martha Anderson (*Office of Strategic Initiatives, Library of Congress*), Steve Knight (*Digital Strategy Implementation, National Library of New Zealand*), Natascha Schumann (*NESTOR, German Network of Expertise in Digital Long-term Preservation*)
- **International Approaches to Web Archiving Panel Discussion** 305
Moderator: Richard Boulderstone (*International Internet Preservation Coalition and The British Library*)
Panelists: Thorsteinn Hallgrímsson (*National and University Library of Iceland*), Birgit N. Henriksen (*The Royal Library, Denmark*), Helen Hockx-Yu (*The British Library*), Gildas Illien (*National Library of France*), Colin Webb (*National Library of Australia*)
- **Training and Curriculum Development Panel Discussion** 307
Moderator: Frances Boyle (*Digital Preservation Coalition*)
Panelists: Nancy McGovern (*Inter-University Consortium for Political and Social Research*), Kevin Ashley (*University of London Computer Centre*), Rachel Frick (*Institute of Museum and Library Services*), Joy Davidson (*Digital Curation Centre*)

Author Index 309

iPRES 2008 Programme and Organising Committee

Programme Committee

Programme Chair: Adam Farquhar, *The British Library*
Frances Boyle, *Digital Preservation Coalition*
Patricia Cruse, *University of California*
Neil Grindley, *Joint Information Systems Committee*
Heike Neuroth, *State and University Library-Goettingen*
Oya Rieger, *Cornell University*
Shigeo Sugimoto, *University of Tsukuba*
Andrew Wilson, *National Archives of Australia*

Organising Committee

Organising Chair: Jane Humphreys, *The British Library*
Doreen Bonas, *The British Library*
Pete Carr, *The British Library*
Alison Faraday, *The British Library*
Carol Jackson, *Digital Preservation Coalition*
Suvi Kankainen, *The British Library*
Rui Miao, *The British Library*
Charlotte Orrell-Jones, *The British Library*
John Overeem, *The British Library*
Anna Spiering, *Leiths Catering*
Lawrence Christensen, *The British Library*
Mark Walton, *The British Library*
Colin White, *The British Library*

Digital Preservation Policy: A subject of no importance?

Neil Beagrie*, Najla Rettberg**, Peter Williams ***

*Charles Beagrie Ltd
www.beagrie.com
neil@beagrie.com

** Charles Beagrie Ltd
www.beagrie.com
najla.rettberg@beagrie.com

*** Charles Beagrie Ltd
www.beagrie.com
peter.williams@beagrie.com

Abstract

There are relatively few digital preservation policies within institutions: is digital preservation a subject of no importance? This paper presents ongoing work and findings from a JISC funded study on institutional digital preservation policies which aims to provide an outline model for digital preservation policies and in particular to analyse the role that digital preservation can play in supporting and delivering key strategies for Higher Education Institutions in areas such as research and teaching and learning. Although focussing on the UK Higher Education sector, the study draws widely on policy and implementations from other sectors and countries and will be of interest to those wishing to develop policy and justify investment in digital preservation within a wide range of institutions.

Introduction

A recent synthesis of the UK Joint Information Systems Committee's digital preservation and records management programme noted that 'the costs and benefits of developing a coherent, managed and sustainable approach to institutional preservation of digital assets remain unexplored' (Pennock, 2008). Across many sectors the development of institutional preservation policies is currently sporadic and digital preservation issues are rarely considered in key strategic plans. The lack of preservation policies and as a result the lack of consideration of digital preservation issues in other institutional strategies is seen as a major stumbling block.

This paper presents the current work and emerging findings of a new JISC -funded study (completing late September 2008 and to be published Autumn 2008) to help institutions the UK Higher Education sector understand, develop and implement relevant digital preservation policies.

Institutions may have a range of central and devolved functions and departments that will need to consider digital preservation in some form. The study is therefore ensuring that it promotes approaches to policy and guidance which will underpin and inform the activities of a wide range of relevant functions and stakeholders within institutions.

The research that has been undertaken in the course of this study references existing institutional policies and

also seeks to include information from outside of the UK HE/FE sector where appropriate. It does not have resources to develop recommendations for all areas from scratch but has referenced and build upon other work, case studies, and tools and services and seeks to identify and position its recommendations to complement existing resources.

Its aim therefore has been to produce a practical "how to" guide for developing an institutional digital preservation policy. It contains strategic policy advice supported by further reading sections which select and provide brief descriptions of key existing resources to assist implementation using specific strategies and tools.

We understand the very different types of institutional needs that need to be supported by the study. We are therefore including guidance on how to tailor a policy for the needs of a specific institution or function. This combined with a modular approach should allow selection and tailoring for a wide range of individual needs.

Finally but perhaps most importantly, we have recognised developing an institutional preservation policy will only be worthwhile if it is linked to core institutional business drivers and strategies: it cannot be effective in splendid isolation. We have therefore devoted significant effort to mapping and linking a preservation strategy to other core university policies including research and teaching and learning.

The format of the remainder of this paper is an overview of progress to date (August 2008) focussing on the development of a model policy and the analysis of high-level institutional strategies from UK universities and potential support for them from digital preservation activities. This is still very much a work in progress and the reader is encouraged to consider the completed version of this work which will be published by JISC in Autumn 2008 and presented at the conference in late September.

Institutional Digital Preservation Policies

After consulting a large range of resources and example policies, it is clear that whilst a high-level policy

framework is needed, a certain degree of practical guidance, to implementation level, must also be offered.

The outline model policy we have created is based on some of the principal themes picked out from a variety of existing digital preservation policies identified and analysed in the desk research. Some key strands are shared in almost all the policies examined: preservation objectives; mission statement; contextual links; financial support; staffing; intellectual property issues. The policy is comprised of two parts, policy and implementation. Policy level is examined in more detail and includes direction on how to structure these high level policy statements and highlights how the principle clauses can tie into other key organizational policies. The implementation level includes technical guidance, containing information about metadata and auditing as well as references to distributed archiving and standards such as the Open Archival Information System (OAIS) Reference Model (CCSDS, 2002). Particular policies/documents of note for our study have been from: the UK Data Archive (Woollard, 2008); the former Arts and Humanities Data Service (James, 2004); the JISC/NPO Beagrie-Greenstein strategic framework for creating and preserving digital resources (Beagrie and Greenstein, 2001); the Interuniversity Consortium for Political and Social Research (McGovern, 2007); the Canadian Heritage Information Network (Canadian Heritage Information Network, 2004); University of Columbia (Columbia University, 2006); and the Cedars Guide to Collection Management (The Cedars Project, 2002). While the research focussed on policies from Higher and Further Education, the British Library (British Library, nd) and the UK National Archives (Brown, 2003) have the most comprehensive technical and administrative strategies. A paper of particular note is the Preserv digital preservation survey report (Hitchcock, Brody, Hey, and Carr, 2007).

While the JISC 04/04 digital preservation programme projects (Pennock, 2008) were varied in their outcomes, many of the results can be synthesised and drawn into the report. Tools are hard to review as it is not yet fully examined how they are received or used within the community. We have thus had to be selective as to what tools are pointed to in the study. With regard to standards, RLG/NARA's Trustworthy Repositories Audit & Certification: Criteria and Checklist (TRAC) (RLG/NARA, 2007) is very comprehensive and is often cited, along with the OAIS reference model as a key standard.

Function-specific areas were looked at such as: e-journals, IRs, organisational electronic records, digitised images, and research data. There is certainly a commonality between different communities and the similar materials they are preserving, for example policies tend to be clearer and more focussed if homogenous sets of materials are the target contents of a repository. However, clear parallels can be drawn

between a different range of policies, and on the whole they don't differ hugely between these functional areas. Examples are Glamorgan Research Repository (University of Glamorgan, 2008), Jorum (Stevenson, 2006), Ethos (Key Perspectives, 2006), and Preserv (Hitchcock, Brody, Hey, and Carr, 2007).

Looking at the high-level policy objectives in the outline model policy, there is significant scope for mapping over to broader organisational policies, such as Teaching/learning, Research, and ICT/Information strategies.

Other High-Level Institutional Strategies

Universities selected for sampling of their high-level strategies were a mixture of teaching-led and research-led institutions (the latter from the Russell Group) and a Further Education college.

University research and learning and teaching strategies have been looked at in most detail so far. These are discussed below.

Research strategies

The strategies were varied in approach and detail so that it was difficult both to condense their key points into only a few categories and to compare them. In fact, in broad terms the teaching universities were surprisingly similar to those leading in research. The principal shared research strategy aims are to:

- *Maximise staff and research excellence:* and increasing active research staff numbers. Central to these strategies are staff development and support.
- *Provide a high level of administrative support:* Strategy aims include offering a co-ordinated administrative service involving an integrated and expanded Research Office, providing support for staff drafting and submitting applications for research funding, and generally supporting the work of full-time research staff within the Department.
- *Recognise and promote the link between teaching and research:* (this finding applied equally across all university types examined, and was not necessarily stronger in teaching-led universities).
- *Increase income and financial sustainability:* Universities are particularly aware of financial considerations and wish to achieve targets for external research funding, including research student funding, and to invest in institutional funding for selective research activities.

Strategy aims that either only applied to teaching-led universities, or were particularly emphasised were noted. Teaching-led universities tended to:

- Include more peripheral aims such as attracting a diverse student body;

- Place greater emphasis on interdisciplinary and collaborative work, including the involvement of external organisations. It may be that established research-led universities do not need to explicitly state this, whilst teaching universities may see them as an excellent way to raise the research profile;
- Explicitly aim to disseminate research– again, presumably to raise the institutions’ research profiles;
- Offer more staff support for research.

Learning and teaching strategies

The main themes of the learning and teaching strategies of the various universities concern:

- The skills, knowledge and experience of the students;
- The use of e-learning;
- The fostering of excellence through staff development and effective leadership;
- Equality awareness.

Strategies tend to emphasise the development of a wide range of skills. In addition to those related to specific disciplines, others included intellectual, generic, and social skills. These are designed to be transferable, to help foster independent and lifelong learning and ‘the appropriate attitudes and values associated with successful graduates’ (Open University Learning and Teaching Strategy 2004-2008). Teaching and learning aims that promote the employability of students are also, unsurprisingly, common.

There is also an emphasis on students developing research skills, and for teaching to be informed by research.

There is a universal commitment to working with and developing new technologies, including virtual learning environments, e-learning programmes and resources such as access to datasets using powerful search tools and services supported by Library and Learning Resources. Many institutions wish to establish e-learning as an integral part of teaching and learning activities.

Strategies also concern the development and refinement of teaching methods, staff development and the general promoting of the institution through the excellence of its teaching programmes. Some institutions mention developing an effective and enabling educational leadership and management structure in order to facilitate this.

Equality awareness and opportunity are also common themes as are the aims emphasising the need to attract international students as well as those from diverse domestic backgrounds.

Additional High-Level Strategies

The comparison and aggregation of publication schemes provides some useful input on records management but has fewer digital preservation implications at this stage. More recently a selection of university Information or IT strategies, Library and “Special Collection” strategies, and records management have been compared and aggregated and digital preservation impacts are now being assessed.

Conclusions

Overall there were some significant common aspects of the other high-level institutional strategies examined that have important implications for digital preservation and that can be linked into our work on developing institutional digital preservation strategies. These cross-correlations are now being made by the study team. Our work to date reinforces our initial view that for institutions digital preservation must be seen as “a means to an end” rather than an end in itself: any digital preservation policy must be framed in terms of the key business drivers and strategies of the institution.

References

- Beagrie, N and Greenstein, D (2001). *A strategic policy framework for creating and preserving digital collections*. JISC/NPO eLib Supporting Study P3, Library Information Technology Centre South Bank University 1998. Retrieved 20 June 2008 from: <http://www.ukoln.ac.uk/services/papers/bl/framework/framework.html>
- British Library (nd) *British Library Digital Preservation Strategy*. Retrieved 20 June 2008 from: <http://www.bl.uk/aboutus/stratpolprog/ccare/introduction/digitalpresstrat.pdf>
- Brown, A (2003) *Preserving the Digital Heritage: building a digital archive for UK Government records*. The National Archives, Kew, Richmond, UK. Retrieved 14 August 2008 from: <http://www.nationalarchives.gov.uk/documents/brown.pdf>
- Canadian Heritage Information Network (2004). *Digital preservation for Museums: Recommendations: A possible checklist for creating preservation policy*. Retrieved 20 June 2008 from: http://www.chin.gc.ca/English/Digital_Content/Preservation_Recommendations/index.html
- CCSDS (2002). *Reference Model for an Open Archival Information System (OAIS)*. CCSDS 650.0-B-1 Blue Book Issue 1. Retrieved 14 August 2008 from: <http://public.ccsds.org/publications/archive/650x0b1.pdf>
- The Cedars Project (2002). *Cedars Guide to Collection management*. Retrieved 20 June 2008 from: http://www.leeds.ac.uk/cedars/guideto/collmanagement/guideto_colman.pdf
- Columbia University (2006). *Columbia University Libraries: Policy for Preservation of Digital Resources*. Retrieved 20 June

2008 from:
<http://www.columbia.edu/cu/lweb/services/preservation/dlpolicy.html>

Hitchcock, S., Brody, T., Hey, J., and Carr, L. (2007). *Laying the foundations for Repository Preservation Services*. Retrieved 20 June 2008 from:
<http://www.jisc.ac.uk/media/documents/programmes/preservation/preserv-final-report1.0.pdf>

James, H (2004). *Collections Preservation Policy*. Arts and Humanities Data Service, King's College London. Retrieved 20 June 2008 from:
<http://www.ahds.ac.uk/documents/colls-policy-preservation-v1.pdf>

Key Perspectives Ltd and UCL Library Services (2006). *Evaluation of options for a UK Electronic Thesis Service*. Retrieved 20 June 2008 from:
<http://www.keyperspectives.co.uk/openaccessarchive/reports/EThOS%20Report%20-%20final%20published%20version.pdf>

McGovern, N (2007). *ICPSR Digital preservation policy framework*. Retrieved 20 June 2008 from:
<http://www.icpsr.umich.edu/DP/policies/dpp-framework.html>

Pennock, M (2008). *JISC Programme Synthesis Study: Supporting Digital Preservation & Asset Management in Institutions*. Joint Information Systems Committee (JISC). Retrieved 14 August 2008 from:
http://www.jisc.ac.uk/media/documents/programmes/preservation/404publicreport_2008.pdf

RLG/NARA (2007). *Trustworthy repositories Audit and Certification: Criteria and Checklist*. OCLC and Centre for Research Libraries. Retrieved 20 June 2008 from:
<http://www.crl.edu/PDF/trac.pdf>

Stevenson, J (2006). *Jorum preservation watch report*. Retrieved 20 June 2008 from:
http://www.jorum.ac.uk/docs/pdf/Jorum_Preservation_Watch_Report.pdf

University of Glamorgan (2008). *University of Glamorgan Online Research Repository Policy*. Retrieved 20 June 2008 from: <http://lcss.glam.ac.uk/gor/options/>

Woollard, M (2008). *UK Data Archive Preservation policy*. Retrieved 17 June 2008 from: <http://www.data-archive.ac.uk/news/publications/UKDAPreservationPolicy0308.pdf>

Modeling Organizational Preservation Goals to Guide Digital Preservation

Angela Dappert, Adam Farquhar

The British Library
Boston Spa, Wetherby, West Yorkshire LS23 7BQ, UK
angela.dappert@bl.uk, adam.farquhar@bl.uk

Abstract

Digital preservation activities can only succeed if they go beyond the technical properties of digital objects. They must consider the strategy, policy, goals, and constraints of the institution that undertakes them and take into account the cultural and institutional framework in which data, documents and records are preserved. Furthermore, because organizations differ in many ways, a one-size-fits-all approach cannot be appropriate.

Fortunately, organizations involved in digital preservation have created documents describing their policies, strategies, workflows, plans, and goals to provide guidance. They also have skilled staff who are aware of sometimes unwritten considerations.

Within Planets [Farquhar 2007], a four-year project co-funded by the European Union to address core digital preservation challenges, we have analyzed preservation guiding documents and interviewed staff from libraries, archives, and data centers that are actively engaged in digital preservation. This paper introduces a conceptual model for expressing the core concepts and requirements that appear in preservation guiding documents. It defines a specific vocabulary that institutions can reuse for expressing their own policies and strategies. In addition to providing a conceptual framework, the model and vocabulary support automated preservation planning tools through an XML representation.

Introduction

This paper introduces a conceptual model and vocabulary for preservation guiding documents. Preservation guiding documents include documents, in a broad sense, which specify requirements that make the institution's values or constraints explicit and influence the preservation planning process. They may be policy, strategy, or business documents, applicable legislation, guidelines, rules, or even a choice of temporary runtime parameters. They may be oral representations as well as written representations in databases, source code, web sites, etc..

The model and vocabulary can be shared and exchanged by software applications. They offer a starting point for creating individualized models for an institution. Below, we show how they can be used to describe requirements for individual institutions, possibly, but not necessarily, in a machine-interpretable form. Furthermore, we show how these requirements can then be used in the context of comprehensive preservation planning.

To perform the analysis, the team used a combination of top-down and bottom-up methods. We examined the

literature [e.g. ERPA 2003, Solinet 2008, ALA 2007, JISC 2006, PADI 2008, Cornell 2008, CRL 2008] to create a top-down model from first principles. To complement this, we analyzed actual preservation guiding documents of archives, national libraries, and data centers for their content [e.g. Australia 2002, Hampshire, Georgia 2005, UKDA 2008, Florida 2007], and interviewed decision makers [Dappert 2008] to determine factors that influence their preservation choices. We extracted relevant concepts and vocabulary from the material to populate our model and compiled a list of example requirements. A more detailed description of this work can be found in [Dappert 2008]. Aspects of this model were based on or developed together with ideas in the TNA conceptual model which underlies PRONOM [Sharpe 2006], the PLANETS conceptual model [Sharpe 2008], and the OAIS model [CCSDS 2002].

Context

The context of our conceptual model is the process of preservation planning for a digital collection [Strodl 2006]. The goals of this process are to

- identify which parts of the collection present the greatest risks.
- identify candidate preservation actions that could be taken to mitigate the risks.
- evaluate the candidate preservation actions to determine their potential costs and benefits. The cost includes the cost of executing the action, the cost of needed infrastructure for sustaining the results of the action, and the cost of essential characteristics lost in the action (e.g. loss of authenticity) etc.. The benefits come from mitigating the risks and increase in proportion to value of the object and the severity of the risk. The costs and benefits are not necessarily monetary.
- provide justified recommendations for which actions to execute on which collections.

All of these activities should be based on institutional requirements which extend beyond considering file formats and characteristics of individual digital objects to take into account the goals and limitations of the institution, features of its user community, and the environment in which its users access digital content.

The Core Conceptual Model

The core conceptual model implicitly describes the institution and consists of the components in Figure 1. In

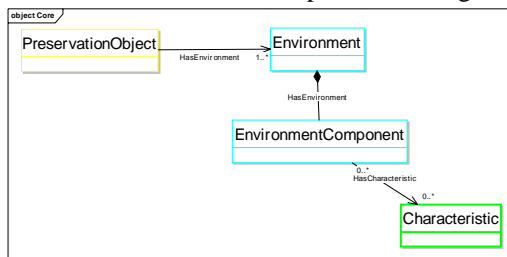


Figure 1 Institutional Data Model

summary, any Preservation Object has one or more Environments. Every Environment in which the Preservation Object is embedded consists of one or more Environment Components, such as hardware and software components, the legal system, and other internal and external factors. Environment Components are described through their Characteristics, which are Property / Value pairs. We realized early that requirements express constraints on many levels of granularity. We, therefore, defined **Preservation Objects** as follows:

A Preservation Object is any object that is directly or indirectly at risk and needs to be digitally preserved.

and introduced the following **Preservation Object Types** as illustrated in Figure 2:

Collection, Deliverable Unit, Expression, Component, Manifestation, Bytestream.

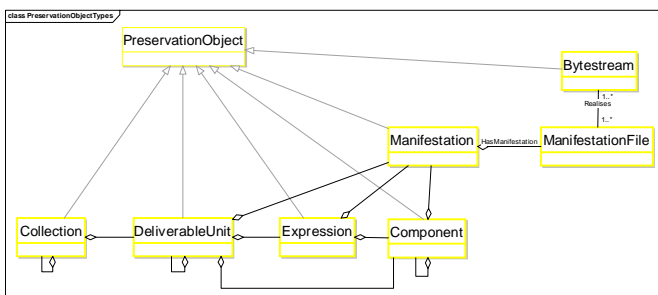


Figure 2 Preservation Object Types

Each Preservation Object Type is related to an other with the “containedIn” relationship (except that a Bytestream is contained in its Manifestation via its Manifestation File).

A **Bytestream** is the primary, physical Preservation Object. If it is at risk of decay or obsolescence it becomes the object of preservation. We create and execute preservation plans to preserve it. A Bytestream is, however, embedded in a larger context.

A **Manifestation** is the collection of all Manifestation Files that are needed to create one rendition of a logical data object. A Bytestream is realised by its **Manifestation File**. Manifestation and Manifestation File are logical descriptions of physical Bytestreams.

Collection, Deliverable Unit, Expression, and Component are logical objects.

In the simplest case, a Bytestream, Manifestation File, and Manifestation have a one-to-one correspondence.

For example, a book that is represented as a single PDF file in the PDF format.

In other cases, however, several Bytestreams may be contained in one Manifestation File and several Manifestation Files may be contained in one Manifestation. For example, several image Bytestreams might be contained in a single Manifestation File.

Example:

- A digital file (Bytestream) is part of its Manifestation (e.g. a MPEG-4 video Bytestream is part of an HTML Manifestation of an article).
- This Manifestation represents an Expression of this article, the specific intellectual or artistic form that the article takes as it is realized, which contains a video stream. There may be other Expressions, such as a static still image Expression that holds an image in place of the video stream.
- All Expressions of this article make up the Deliverable Unit. The Deliverable Unit is the abstract concept representing the distinct intellectual creation, which is the article. There might be several Expressions with several Manifestations of the same article (e.g. an HTML, a PDF, an XML, a publisher specific format).
- The article is part of another Deliverable Unit, the issue (hence the recursive link in the diagram).
- And the issue is part of the Deliverable Unit journal, which is the logical object describing all issues of the same title.
- The journal belongs to a Collection. The Collection might be static for the institution, such as the Science Collection, or it might be determined dynamically, such as the Collection of all articles that contain TIFF3.0 files. Collections may contain digital and non-digital objects.
- Collections may be recursively contained in larger Collections.
- Finally, all Collections are part of the whole institution, which is modelled as the top-level Collection.
- Deliverable Units or Expressions consist of logical Components for which Values for Characteristics can be measured or assigned, such as a “table” Component or a “title” Component of a journal article.

Since higher-level objects (such as the Manifestation that includes the affected Bytestream, and the Collection in which this Manifestation is held) are indirectly affected by its preservation need, they also need to be considered during preservation planning. Thus, they are indirectly Preservation Objects. Conversely, an institution can not consider the preservation of each individual data object in isolation. Institutions need to take a global look at all their Collections and resources in order to prioritise their Preservation Actions and co-ordinate preservation activity. In order to facilitate this, the model goes well beyond planning for the individual data object.

Every Preservation Object has one or more **Environments** which may fulfil different roles. For example, a Bytestream or a Manifestation may have creation, ingest, preservation, and access Environments; a Collection may have an internal, a physical delivery, and an online delivery Environment.

Environments for Preservation Objects at a higher level must accommodate the requirements of Preservation Objects at a lower level. As long as a ByteStream is part of its Manifestation, it will live in the Manifestation's Environment. When it is taken out of the Manifestation's Environment, for example to be used in a migration, then the ByteStream's individual Environment requirements will influence the Environment of its new Manifestation. It is worth noting that it may not be possible to derive the best Environment from a ByteStream's file format. If, for example, a Word file contains only text without formatting, headers and tables, etc., then a .txt output might be considered perfectly adequate, even though this would in general not be considered an ideal migration format for a Word file. Institutions may wish to specify whether an Environment is necessary, recommended, or acceptable. Every Environment consists of a number of Environment Components. These include the commonly considered software and hardware environments. They also include factors such as the community, legal or budgetary restrictions. **Environment Components** are defined as follows:

A factor which constrains a Preservation Object and that is necessary to interpret it.

There is a close relationship between an Environment and an extended notion of Representation Information as it is defined in OAIS [CCSDS 2002]. Other examples of extended notions of Representation Information are discussed in [Brown 2008].

The top-level **Environment Component Types** (see Figure 4) include software, hardware, community and Content/Self. The name 'Content/Self' refers to the intellectual content of the Preservation Object. In the case of Preservation Objects which are individual items, the word 'content' or 'intellectual content' provides a good name, but in the case of Preservation Objects which are collective items the word 'self' better reflects the intention. The Content/Self has associated three factors:

- its semantic and syntactic interpretation,
- the format in which it is encoded, and
- its physical realisation.

The Content/Self is actually an Environment Component; several may be associated with a single Preservation Object. They can then be treated like other Environment Components with their associated Characteristics and Values and be used in the preservation planning process in a uniform way. We decompose the OAIS "Digital object" into two aspects: the intellectual content Content/Self and its physical Realisation.

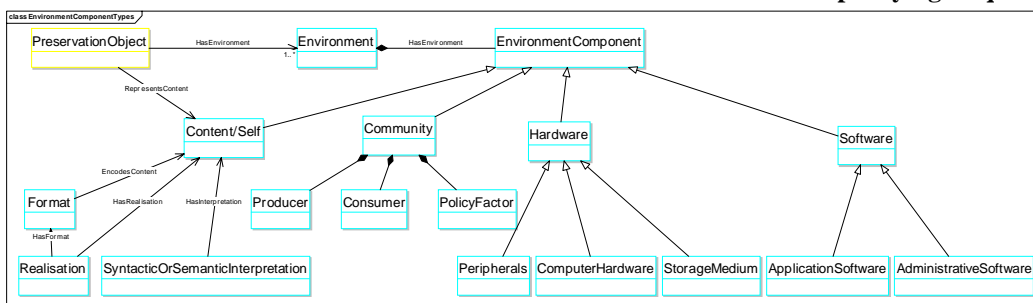


Figure 4 Environment Component Types

See the full report [Dappert 2008] for additional Environment Component Types that have been extracted from preservation guiding documents. Policy Factors, in particular, are discussed in depth.

Characteristics describe the state of Environment Components as Property / Value pairs. Values may be stored directly as object values, referenced indirectly through registries or in inventories, or extracted dynamically through characterisation processes. The vocabulary for Properties can be found in the full report [Dappert 2008].

The Full Conceptual Model

The full conceptual model which describes the institution embedded in the preservation planning domain consists of the components in Figure 3.

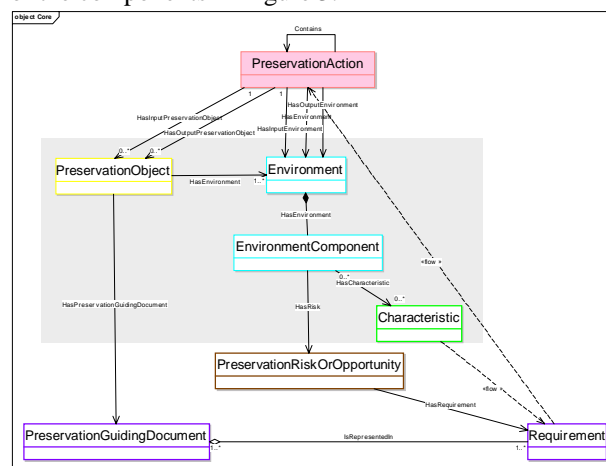


Figure 3 Full Data Model. The Shaded Area Indicates the Core Model.

Degradation of Preservation Objects is caused by two things:

- Preservation Risks
- Executing imperfect, lossy Preservation Actions

Acceptable levels of degradation are defined in an institution's Requirements, which specify permissible or desirable Characteristics of Environment Components. They make the institution's values explicit, influence the preservation process, and are captured in Preservation Guiding Documents.

Changes to an Environment Component, such as obsolescence of hardware or software components, decay of data carriers, or changes to the legal framework may introduce **Preservation Risks**.

An individual institution's Preservation Risks are specified in **Risk Specifying Requirements**. Whenever Char-

acteristics of a Preservation Object's Environment Component violate the Values which are specified in the Requirement then the Preservation Object is considered at risk. Once a Risk Specifying Requirement is

violated, a preservation monitor-ring process should notice this and trigger the preservation planning process. It, in turn, determines the optimal Preservation Action to mitigate this risk.

Preservation Object Selecting Requirements are a sub-type of Risk Specifying Requirements which specifies which subset of Preservation Objects is at risk.

A composite **Preservation Action** may consist of elementary Preservation Actions and may include conditional branches and other control-flow constructs.

When a Preservation Action is applied to a Preservation Object and its Environment, it produces a new Preservation Object and/or a new Environment in which the Preservation Risk has been mitigated. Every Preservation Action, therefore, has not only an Input Preservation Object and (at least one) Input Environment, but also an Output Preservation Object and Output Environment. For example, if a Microsoft Word Bytestream is migrated to a PDF Bytestream this results in a new Preservation Object, which might have slightly different Characteristics, but also a new Environment in which it can be used – in this case the platform needs to at least contain a PDF viewer. This approach works for migration, emulation, hardware and other solutions.

For any given Preservation Object and its Environment, there are multiple possible Preservation Actions which might mitigate the Preservation Risk. Which of these Preservation Actions is the most suitable for the Preservation Object can be derived from the information in the **Requirements**.

In order to determine whether an abstract Requirement is applicable and satisfied, one needs to evaluate the concrete Values of the Characteristics of Environment Components which describe the actual Preservation Objects or the concrete Values of a candidate Preservation Action at a given time.

Machine-interpretable Requirements can be expressed in OCL (the Object Constraint Language). They refer solely to concepts and vocabulary contained in the model. Requirements may define the context, pre- and post-conditions, have associated Importance Factors, which specify the importance of the requirement for the institution, as well as a specification of the Operators to be applied to determine whether the requirement is satisfied, and a Tolerance which specifies to what degree deviation from the Requirement can be tolerated.

Requirement Types

During our literature and document analysis, we extracted Requirements that we categorized into the Requirement Types depicted in Figure 5. Besides Risk Specifying Requirements, which were already discussed earlier, there are further Requirement Types.

Preservation Guiding Requirements specify which kinds of Preservation Actions are desirable for the Preservation Object. For example: The size of the

Preservation Action’s output Preservation Object should not exceed a maximal size as set by the institution. They are dependent on

- which input Characteristics of the Preservation Objects need to be met to consider the Preservation Action.
- which output Characteristics of the Preservation Objects are permissible or desirable (either in absolute terms or in relationship to Characteristics of the input Preservation Object, which might be a derivative or the original submitted to the institution).
- which Characteristics of the Preservation Action itself are desirable.

Action Defining Requirements (sub-type of Preservation Guiding Requirement) define which kinds of Preservation Actions are desirable independent of the Characteristics of the Preservation Object, but dependent only on the Characteristics of the Preservation Action itself. For example PDF may, for a given institution, not be an acceptable preservation output format of a Preservation Action (independent of any input Characteristics of Preservation Objects).

Significant Properties (sub-type of Preservation Guiding Requirement) are often limited to Characteristics of Bytestreams or Components for which it is possible to evaluate Values automatically. Our definition is close to the more expansive one expressed by Andrew Wilson, National Archives of Australia: “the Characteristics of digital objects that must be preserved over time in order to ensure the continued accessibility, usability, and meaning of the objects, and their capacity to be accepted as evidence of what they purport to record.” We, too, consider Significant Properties at any level of Preservation Object Type. We, however, treat them as Requirements rather than Characteristics. While Preservation Guiding Requirements in general can combine constraints on multiple Characteristics on several levels of Preservation Object Types, Significant Properties refer to one Characteristic at a time.

Preservation Process Guiding Requirements (sub-type of Preservation Requirement) describe the preservation process itself independent of the Characteristics of the Preservation Object or the Preservation Actions. For example: A preservation planning process should be executed for every data object at least every 5 years,

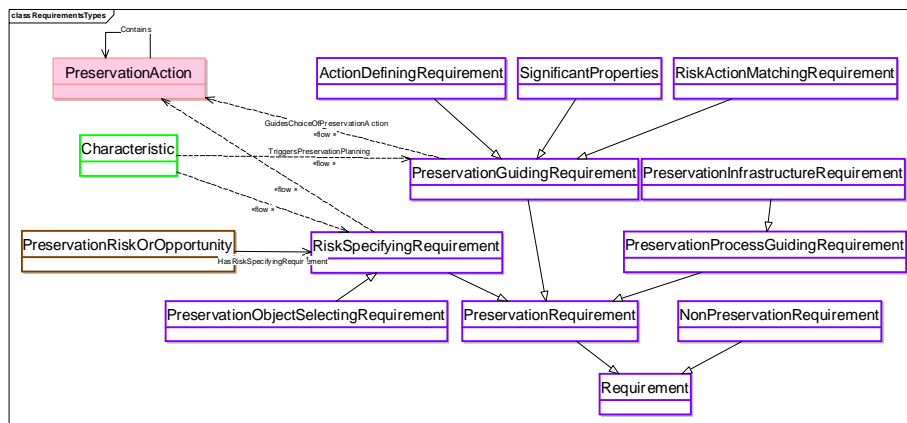


Figure 5 Requirements Types

independent of the Preservation Risks that are established for this data object. These requirements do not influence the preservation planning process.

Preservation Infrastructure Requirements (sub-type of Preservation Process Guiding Requirement) are particularly prominent in preservation guiding documents. They specify required infrastructure Characteristics with respect to security, networking, connectivity, storage, etc.. For example: Mirror versions of on-site systems must be provided.

Non-Preservation Requirements (sub-type of Requirement) specify the set of requirements found which specify processes relevant to preservation, but not part of preservation itself.

Risk / Action Matching Requirements (sub-type of Preservation Guiding Requirement) specify that a candidate Preservation Action has to be an appropriate match to a given Preservation Risk. They are rarely stated explicitly in preservation guiding documents.

Preservation Risk Types are (see Figure 6)

- **NewVersion:** A new version of the Environment Component is available. This creates a risk of future obsolescence, or a risk of having to support too many versions of this Environment Component.
- **NotSupportedOrObsoleteSupport:** The Environment Component is no longer sufficiently supported. This creates a risk that support will cease altogether, rendering the Environment Component non-functional.
- **DeteriorationOrLoss:** The Environment Component is deteriorating or has been lost. Reconstruction or replacement become necessary.
- **Proprietary:** The Environment Component is proprietary. There is a risk that it cannot be replaced since the specifications for it are unknown.
- **UnmanagedGrowth:** The institution's Environment is becoming too diverse to manage. A normalization Preservation Action is needed to simplify or unify the

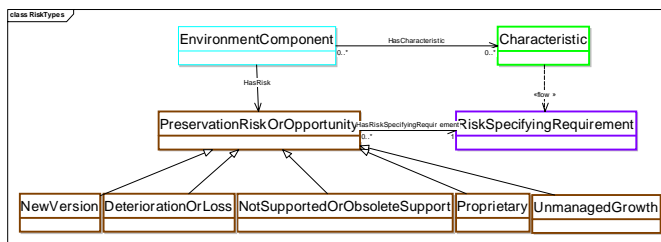


Figure 6 Risk Types

Corresponding to every Preservation Risk Type and the type of the affected Environment Component and Preservation Object, there are appropriate

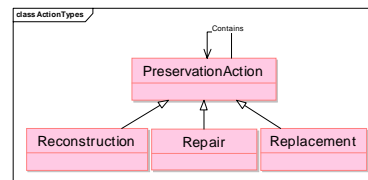


Figure 7 Action Types

Preservation Actions. For example, the risk of data carrier failure can be mitigated by a carrier refresh. The risk of file format obsolescence can be mitigated by migrating objects to an alternative format.

Preservation Action Types are replacement, repair and reconstruction (See Figure 7).

The diagram (Figure 8) and table (Figure 9) illustrate the correspondence between Preservation Risk Type, Environment Component Type, Preservation Object Type and Preservation Action Type.

Most of them are self-explanatory. Some deserve some comment:

- Modification of Content/Self might represent an action such as the reconstruction of a deteriorated file, or a file that is modified in order to satisfy new legal requirements.
- One possible Preservation Action is to not do anything (wait and see).

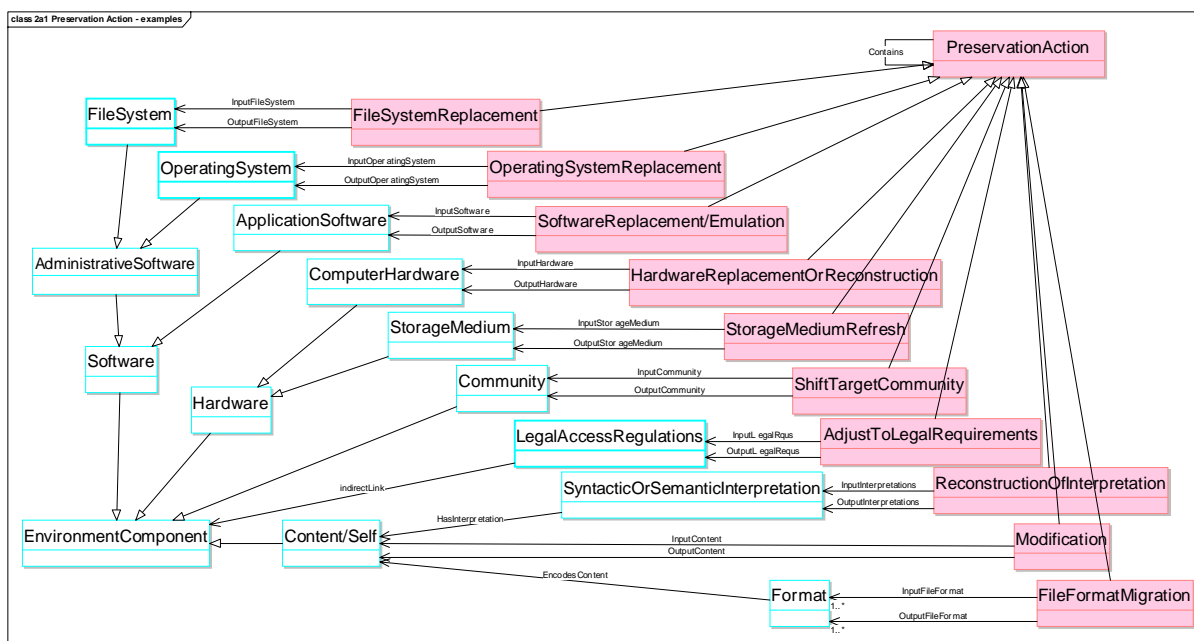


Figure 8 Risk-Action Matching Requirement

Environment.

Example Risks	Preservation Object Type	Environment Component Type	Preservation Risk Type	Preservation Action Type
Data carriers deteriorate and cannot be read	Bytestream	Data Carrier	Deterioration	Replacement
The data object becomes corrupted on the carrier and the original byte stream cannot be retrieved.	Bytestream	Realization	Deterioration	Reconstruction
Essential hardware components are no longer supported or available	Collection	Hardware	Not supported	Replacement
Software components are proprietary and the dependence is unacceptable to the institution.	Collection	Software	Proprietary	Replacement
The community requires new patterns of access, such as access on a mobile phone, rather than a workstation	Collection	Hardware and Software	Obsolete	Replacement
File formats become obsolete.	Bytestream	Format	Obsolete	Replacement
The legislative framework changes and the data or access to it has to be adapted to the new regulations	Collection	Legislation	New Version	Replacement

Figure 9 Risk-Action Matching Requirement

- Migration does not always imply that a different file format is chosen. For example, a collection might contain PDF files which do not include all of the fonts needed. One might migrate them from PDF (without embedded fonts) to PDF (with all fonts embedded).
- The needs of the target community might be a deciding factor for the choice of Preservation Actions, and, conversely, the choice of Preservation Actions will shape and change the community, just as it changes other Environment Components. Shifting the target community might be a somewhat unintuitive Preservation Action, which is parallel to all other forms of Environment replacement. An example might be turning a research data centre into a history-of-science repository, as the material contained in the collection ceases to live up to contemporary standards of scientific use.
- Community has producers and consumers which may be technical (e.g. repository or IT staff, publishing staff) or content oriented (authors or readers). They may consider a digital object obsolete under different circumstances.

Use to Model Institutional Requirements

The diagram in Figure 10 gives an overview of how the model described in this report can be used to create an institutional preservation guiding document. It introduces the General Model that consists of the concepts and vocabulary that are described in this paper, and the Instantiated Model that an institution might create to reflect its individual state and requirements.

The numbering in the text refers to components in the diagram. Numbering including the letter “a” describes components in the general model. Numbering including the letter “b” describes components in an instantiated model. (1a) The conceptual model, as discussed in this paper, defines

the basic concepts that are needed in the domain of organizational preservation guiding documents and the relationships between them. They comprise Preservation Objects, Environments, Environment Components, Characteristics, Preservation Actions, Risks and Requirements.

(2a) The specific vocabulary defines

- subtypes of the basic concepts,
- properties for all types of Environment Components,
- allowable values for these properties.

It is a representative (i.e. not exhaustive) specific vocabulary.

(3a) The requirements base describes sets of organizational requirements which may be contained in preservation guiding documents. They are expressed solely in terms of the concepts and attributes of our conceptual model and of the specific vocabulary. They may be parameterized so that they can be instantiated to a specific institution’s conditions. We plan to represent requirements in OCL.

(4a) The elements in the conceptual model, the specific vocabulary, and the requirements base can be translated into several implementation specific machine-interpretable representations, for example based on an XML schema.

(1b) The institution chooses which of these concepts are supported in its setting and are needed by its preservation planning service. Since the conceptual model is very concise, in most cases all of the concepts would be expected to be used.

(2b) The institution chooses which specific vocabulary applies to it. The institution also assigns values to the

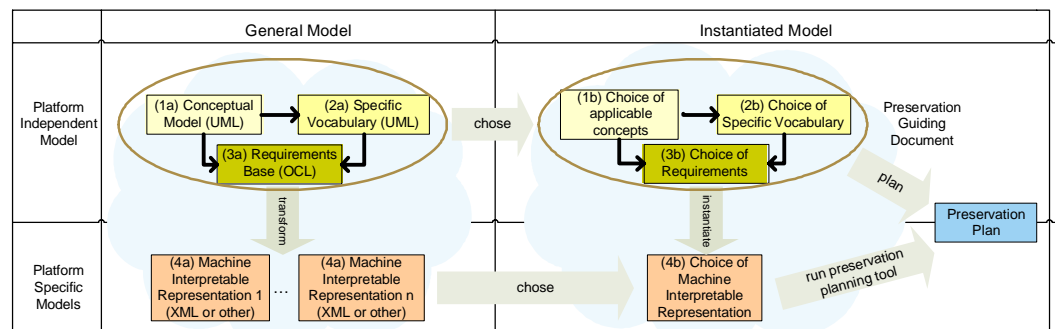


Figure 10 Modelling institutional requirements

Characteristics of its Environment Components if these values will not be measured automatically, or otherwise specifies the method of obtaining measurements or derivations. It will, for example, need registries of tools, formats, and legislative requirements, and need inventories of its collections, software licenses and staff members.

(3b) The institution chooses which Requirements in the Requirements base apply and instantiates them, so that they are now un-parameterized. It specifies Importance Factors, Operators, and Tolerances.

The outputs of steps (1b), (2b) and (3b) form the core part of a preservation guiding document.

(4b) From the choices of steps (1b), (2b), (3b), and the choice of machine-interpretable language results an instantiated machine-interpretable description of the institutional Requirements. This serves as a basis for automated preservation planning. Many requirements in preservation guiding documents, especially on higher institutional levels, may not be machine-interpretable, but it can still be useful to represent the machine-interpretable subset for automatic evaluation.

The planning tool now matches the Requirements in the machine-interpretable version of the preservation guiding document (4b) against the state of the institution to see which Preservation Actions can best satisfy the Requirements under the given state.

Use to Perform Comprehensive Preservation Planning

This model is well-suited for describing any Preservation Object Type and a wide range of preservation processes (e.g., monitoring, planning, characterisation).

First, for example, characterisation tools are defined to work on the Component and Bytestream level. But there are also tools that characterise on a higher level, such as collection profiling tools which analyse Characteristics of a Collection at a given time and produce profiles describing the Collection. They could in principle share the conceptual model and associated processes.

Second, preservation planning needs to compare the Characteristics of a Preservation Object before and after the execution of a candidate Preservation Action in order to evaluate the action against an institution's Requirements. The result is an evaluation score for how suitable each candidate Preservation Action is with respect to the Institution's Requirements. The utility analysis of Plato [Becker 2008] is an example of this.

Preservation Requirements express constraints on all levels of Preservation Objects in the Preservation Object hierarchy (e.g. budgetary constraints on the Collection level; preserving interactivity at the Expression level) and might even mix Characteristics from several levels (e.g. specifying constraints on Collections which contain Bytestreams with a certain Characteristic).

Since each possible Preservation Action may impact multiple levels in the Preservation Object hierarchy, the evaluation of a Preservation Action must be determined on all levels. That is, for every candidate Action, we can evaluate how well it satisfies the Requirements associated with a specific Bytestream, as well as how well it

satisfies the Requirements for the whole of its Manifestation, Deliverable Unit, or even Collection.

If for example, a concrete Preservation Action exceeds the Institution's budget, then it need not be considered for a given Bytestream. Equally, if it violates a Collection principle, even though it would be very suitable for preserving a specific Manifestation, it need not be considered. This sort of higher-level constraint is very useful to rule out unsuitable candidate Preservation Actions at a lower level.

Conversely, it is necessary to not just evaluate a concrete Preservation Action's utility in isolation on a lower level, but rather place it in a higher level context. When combining the evaluations from lower levels, with constraints on the higher level, then the evaluation of a Preservation Action might shift in the more global perspective. Planning algorithms need to take this into account.

For example,

- Preservation Action A is considered more suitable than Preservation Action B in the evaluation for a digital file. But if we look now onto a higher level then it might not be possible to combine Preservation Action A with the suggested Preservation Actions for the other files in the Manifestation, which is an inherent Preservation Process Guiding Requirement on Manifestation level. This might, for example, be the case if the Actions' outputs require incompatible environments.
- For a .png file we decide that it is best migrated to a .gif file. When we look at the enclosing Deliverable Unit "web page" we see that the references to the image are broken and that the best Action would now add the Preservation Action "rename the links". When we look at the next higher Deliverable Unit "website" we see that they use java script for their links. The renamed links would not work. The best option is now to use a redirect list for the web server to the image on the server side instead of adding the Preservation Action "rename the links".

It is necessary for the Environment at a higher level to accommodate the Environments required at a lower level. For example, the Manifestation Environment needs to accommodate the Environment for all files in the Manifestation.

Conclusion

This paper introduced a conceptual model and vocabulary for preservation guiding documents. We showed how the model and vocabulary can be used to model requirements for individual institutions, possibly in a machine-interpretable form, and how these requirements can then be used to perform *comprehensive* preservation planning that

- accommodates a full range of preservation planning processes such as monitoring, characterization, comparison of characteristics, and evaluation of candidate preservation actions.
- allows processes to be associated with a full range of entities from institutions, and collections, down to byte-streams and atomic logical components of digital

objects. It is, for example, necessary to refer to characteristics at a lower level to represent requirements at a higher level. For example, in order to specify “collections which contain files that exceed 1 GB”, you need to be able to specify the file property “file size” as well as collection properties.

- considers technical as well as organizational properties. Some institutions mandate a particular “technical preservation strategy” (migration, for example) at the preservation policy level, regardless of the lower level technical requirements. This demonstrates the need to integrate institutional and data object considerations in the conceptual model.
- accommodates all types of preservation actions, from software actions (e.g. migration, emulation, file repair), hardware related actions (e.g. data carrier replacement or hardware replacement / reconstruction / repair), to organisational actions (e.g. adapt processes to new legislation, adapt to new requirements of the designated community).

The conceptual model presents a simple but expressive representation of the preservation planning domain. The model and vocabulary can be shared and exchanged by software applications. They offer a convenient starting point for creating individualized models for an institution; this holds true even if the institution does not require a machine-interpretable specification. The model views preservation planning as a process that identifies and mitigates risks to current and future access to digital objects.

This paper represents the current state of our work. We expect to modify and improve it over the coming year in response to feedback and experience applying it in the Planets project.

Acknowledgements

The authors would like to thank colleagues on the Planets project, in particular Bart Ballaux, Michaela Mayr, Rob Sharpe, Sara van Bussel for contributions to this work.

Work presented in this paper is partially supported by the European Community under the Information Society Technologies (IST) Programme of the 6th FP for RTD - Project IST-033789. The authors are solely responsible for the content of this paper.

References

Farquhar, A., and Hockx-Yu, H. *Planets: Integrated services for digital preservation*. Int. Journal of Digital Curation 2, 2 (November 2007), 88–99.

ERPA, Electronic Resource Preservation And Access Network Sept. 2003. *ERPA Guidance: Digital Preservation Policy Tool*. <http://www.erpanet.org/guidance/docs/ERPANETPolicyTool.pdf>

Solinet. *Contents of a Digital Preservation Policy*. <http://www.solinet.net/Preservation/Resources%20and%20Publications/Contents%20of%20a%20Digital%20Preservation%20Policy.aspx>

ALA, American Library Association 2007. *Definitions of Digital Preservation*. ALA Annual Conference, Washington, D.C., June 24, 2007 <http://www.ala.org/ala/alcts/newslinks/digipres/PARSdigdef0408.pdf>

JISC Digital Preservation and Records Management Programme Nov. 2006. *Digital Preservation briefing paper*. http://www.jisc.ac.uk/publications/publications/pub_digipreservationbp.aspx,

PADI *Preservation Access to Digital Information*. <http://www.nla.gov.au/padi/>

Cornell University Library 2008. *Digital Preservation Management: Implementing Short-term Strategies for Long-term Problems. Tutorial*. <http://www.library.cornell.edu/iris/tutorial/dpm/terminology/strategies.html>

CRL, The Center for Research Libraries and OCLC Online Computer Library Center, Inc., Feb 2008. *Trustworthy Repositories Audit & Certification: Criteria and Checklist*. <http://www.crl.edu/PDF/trac.pdf>

National Archives of Australia, Dec 2002. *An Approach to the Preservation of Digital Records*. http://www.naa.gov.au/Images/An-approach-Green-Paper_tcm2-888.pdf

Hampshire Record Office. *Digital Preservation Policy*. <http://www.hants.gov.uk/record-office/policies/digital.html>

Digital Archives of Georgia 2005. *Digital Preservation Policy*. http://sos.georgia.gov/archives/who_are_we/rims/digital_History/policies/policy%20-%20Digital%20Preservation%20Policy.pdf

UKDA, UK Data Archive Mar 2008. *Preservation Policy*

Florida Digital Archive Aug 2007. *Policy Guide*. <http://www.fcla.edu/digitalArchive/pdfs/DigitalArchivePolicyGuide.pdf>

Dappert, A., Ballaux, B., Mayr, M., van Bussel, S., 2008. *Report on policy and strategy models for libraries, archives and data centres*. PLANETS report PP2-D2. <http://www.planets-project.eu/>

Sharpe, R. 2006. *SDB Data Model, V1.R2.M0*. Private correspondence 10-Nov-2006

Sharpe, R. 2008. *PLANETS core conceptual model*, Version 1.4. Private correspondence

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

Strodl, S. et alii, 2006. *The DELOS testbed for choosing a digital preservation strategy*. In Proceedings of the 9th International Conference on Asian Digital Libraries (ICADL'06) (Kyoto, Japan, November 27-30 2006), Springer, pp. 323–332.

Brown, A., 2008. *White Paper: Representation Information Registries*. PLANETS report PC/3-D7 (Restricted to programme participants)

Becker, C., Kulovits, H., Rauber, A., Hofman, H. 2008. *Plato: A Service Oriented Decision Support System for Preservation Planning*. JCDL'08, Pittsburgh, Pennsylvania, USA.

Component Business Model for Digital Repositories: A Framework for Analysis

Raymond J. van Diessen

IBM Global Services
Johan Huizingalaan 765
1066VH, Amsterdam
The Netherlands
Raymond_vanDiessen@nl.ibm.com

Barbara Sierman

National Library of the Netherlands
Prins Willem-Alexanderhof 5
2509 LK The Hague
The Netherlands
Barbara.Sierman@kb.nl

Christopher A. Lee

School of Information and Library Science
University of North Carolina
CB#3360, 100 Manning Hall
Chapel Hill, NC 27599-3360
callee@ils.unc.edu

Abstract

Digital preservation is too big a challenge for any institution or solution supplier to confront on its own. The success of any long-term digital repository will depend upon multiple “open” services provided by a wide range of service providers. No company or organisation in the world is able to provide the preservation solution for all known formats, object types, or policies. Viable approaches are likely to span organizational, institutional and national boundaries. In 2003 the KB, National Library of the Netherlands, in cooperation with IBM, developed the e-Depot as their solution for long-term preservation of digital publications. The core of the e-Depot is IBM’s Digital Information Archiving System (DIAS). This article will discuss the exercise of the KB/IBM Research Group to apply IBM’s Component Business Modelling (CBM) in a digital preservation environment. The CBM map is used by a process called Goal Service Modelling (GSM) to identify candidate services for future versions of the e-Depot. Heat maps are used for impact analysis – to discuss organisational structures, existing hardware and software solutions and business processes in the context of the CBM map. The approach is suggested as a way for other repositories to manage and coordinate their activities, as well complimenting current repository audit and certification activities.

Introduction

There has been a growing professional understanding of what a trustworthy digital preservation repository should look like. The Trustworthy Repositories Audit and Certification: Criteria and Checklist (TRAC 2007), Digital Repository Audit Method Based on Risk Assessment (DRAMBORA) (McHugh et al. 2007), and Network of Expertise in Long-Term Storage of Digital Resources (NESTOR) (Dobratz et al. 2006) all identify measures that institutions should take to support a trustworthy repository, not only with ingest, but also with the other functions, such as preservation planning and access. Tools and services are

now available to support many of the core functions, e.g. the JSTOR/Harvard Object Validation Environment (JHOVE) and Digital Record Object Identification (DROID) for file characterization; kopal Library for Retrieval and Ingest (koLibRI) and Producer - Archive Workflow Network (PAWN) for ingest; Automated Obsolescence Notification System (AONS) for preservation planning; Typed Object Model (TOM) for migration; and Dioscuri and the Universal Virtual Computer (UVC) for emulation. Repositories with crucial file format information are being developed, including PRONOM, the Global Digital Format Registry (GDFR), and IBM Preservation Manager. Several international projects - including Preservation and Long-Term Access through Networked Services (PLANETS) and Cultural, Artistic and Scientific knowledge for Preservation, Access and Retrieval (CASPAR) - aim explicitly to deliver tools and systems to support parts of the digital preservation process. We can expect more tools and services to be developed specifically for long-term digital preservation in the years ahead. While these tools and services provide a valuable contribution to managing digital preservation, it is up to the organisations themselves to integrate and manage these tools and services in their digital preservation environments.

The KB, National Library of the Netherlands, manages the e-Depot. At this moment, 11 million digital objects are stored in the e-Depot. Many new types of material are coming, and they will be more complex (such as websites, e-books and compound objects). In 2006 the KB and IBM formed a Research Group to develop a vision of how to integrate external tools and services into the architecture of the e-Depot. In this exercise IBM’s Component Business Modelling (CBM) method played an important role. This method allows an organisation to map its strategies to relevant business components that support the organisation’s objectives and helps to identify the most important business components. Each of these business components – in the case of the KB, most being departments within the library – will need services to reach their goals. The CBM method provides a framework for

viewing and analyzing the organisation as a network of individual components. Once processes and organization are dissected into discrete understandable and manageable components, the unique activities and associated resources, tools and services for each individual business component can be identified. Through the definition of the business components, the responsibilities for the management of the associated resources is clearly specified.

At the KB, the focus has been on business components related to the long-term digital preservation activities of the e-Depot. Both the actual situation and the future plans were input for the Research Group to identify necessary activities and the supporting services to accomplish the digital preservation goals. The Digital Repository CBM Map helps to determine when and where resources should be focused and how external services and solutions can be integrated.

Component Business Model

IBM has developed the CBM approach to help their clients to map business strategy to business components. Business components are the core building blocks of the organisation. A business component identifies a cluster of activities that together implement some set of capabilities which are offered through services. We will differentiate between business services, which can be provided either with or without the support of Information Technology (IT), and IT services, which are provided completely through software. When we use the term “services,” it will imply both business and IT services. Business components can be managed independently, and their business and IT services can be reused across the organisation. CBM allows an organisation to identify its core business components, and understand where there are opportunities to outsource and/or cooperate with 3rd parties. An individual business component contains the activities and associated resources – such as organisational structure, people, skills and technology – to implement specific capabilities (services) needed by the organisation to achieve its goals.

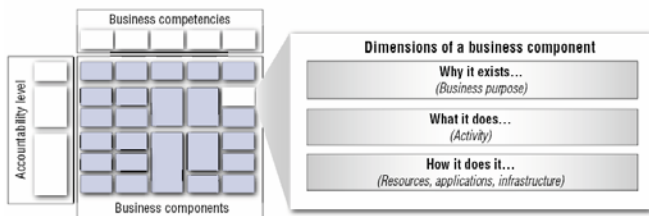


Figure 1: Basic CBM Structure

The description of every business component should involve answering three questions:

- Why it exists – What is the business purpose of

the component?

- What it does – What are the core activities to support the business purpose?
- How it does it - How are the activities to be preformed and what resources are needed, e.g. people and IT support?

The business components are clustered along two dimensions. Horizontally, an accountability level characterizes the scope and intent of activity and decision-making. The three accountability levels used in CBM are Directing, Controlling and Executing:

- **Directing** - strategy, overall direction and policy
- **Controlling** - monitoring, managing exceptions and tactical decision making
- **Executing** - doing the work

Vertically, the major business competencies are identified. Business competencies are large business areas with common global objectives. For example, in a library environment, collection management and customer service management are major business competencies.

KB Objectives

In order to understand the rationale behind the identified business components, we first have to identify the strategic objectives of the KB. The mission statement of the KB identifies four major objectives:

- 1) We give researchers and students access to research information;
- 2) We enable everyone to share in the riches of our cultural heritage;
- 3) We foster the national infrastructure for scientific information;
- 4) We further permanent access to digital information within an international context.

The Research Group focused on the aspects related to Digital Preservation and especially the objective “We further permanent access to digital information within an international context.”

By identifying the key business components needed to support the above objectives of the KB we create a framework for viewing the organisation as a network of individual business components. Once processes and organisation are dissected into discrete understandable and manageable components, the unique activities and associated services can be identified, along with the resources needed to execute them.

The CBM map presented in the next section has been developed with the above organisational objectives in mind. Although the focus of this article is management of digital collections, management of “paper-based” collections will also be supported by the same set of

business components.

Digital Repository CBM Map

Based on the KB’s strategic objectives mission statement discussed above, we identified five major competencies by which to cluster the individual business components:

- **Service Management:** Delivery of collection objects and associated services to the customers of the KB across the supported channels.
- **Collections Management:** Acquisition, processing and cataloguing of all publications, both for the research collection and the deposit collection.
- **Preservation Management:** Facilitating access to the different collections over-time, including addressing media decay and obsolete technology associated with each digital collection.
- **Business Management:** General management of the business of the KB.
- **IT Management:** Management of the overall IT infrastructure.

focussed on IT management can be found in (Ernest and Nisavic 2007).

This article will not elaborate on the more generic components of the Business Management and IT Management competencies. However, we would like to stress that IT Management has some specific objects in relation to the long-term requirements of a digital repository. Not only do the digital collection assets have to be preserved for the long term, but the digital repository solution itself also should be able to adapt to technology changes. This requires the different components of such a digital repository solution to be modular with well-defined interfaces and based on open standards, as well as characterising the preservation environment itself, as it changes over time (Moore 2008).

Figure 2 presents the CBM map created for the KB, with the focus being on the management of digital collections (digital repository). Service Management addresses the services the KB uses to support its customer base. Service Management can be organized along the three CBM

	Service Management	Collection Management	Preservation Management	Business Management	IT Management
Direct	Release Strategy	Collection Strategy	Preservation Strategy		
	Distribution Plan				
Control	Rights Management	Collection Policy Management	Preservation Policy Management		
		Metadata Management	Preservation Planning		
Execute	Reading Room		Delivery & Capture	Characterization	Preservation Action
	Internet	3rd Party	Ingest	Validation	Preservation Research
	Licensing and Royalties Management	Access Management	Collection Storage	Cataloguing	Technology Monitoring
	Packaging & Delivery		Collection Research	Digitalisation	

Figure 2: Digital Repository CBM Map

Service Management, Collection Management and Preservation Management are specific to organisations that manage digital collections, i.e. manage digital repositories. Business Management and IT Management are more generic and needed in any type of organisation. In this article, we have focused on the first three business competencies. A more detailed preservation of a CBM

accountability levels: directing, controlling and executing. Release Strategy and Distribution Plan are part of the general strategy (Direct). Rights Management plays an essential role in determining whether and how content can be delivered to specific customer groups (Control). At the execution level are the different channels through which services are delivered.

The different Service Management business components

are described below in more detail:

- **Release Strategy:** Defines which collections are available to which customer groups, and the collection enrichment services, such as abstracts and classifications, to be associated with each collection or customer group.
- **Distribution Plan:** Defines the specific access strategies to be supported for particular combinations of collections and customer groups.
- **Rights Management:** Controls potential usage restrictions to be enforced on specific collections.
- **Reading Room:** Provides services to the customers in the KB reading rooms.
- **Internet:** Provides remote services to customers via the Internet.
- **3rd Party:** Provides services related to 3rd-party organisations, such as publishers or other cultural heritage institutions.
- **Licensing and Royalty Management:** Manages all licensing, royalty and accounting aspects associated with a particular collection.
- **Access Management:** Provides mechanisms to enforce the particular licenses and rights associated with any given collection, as well as the identification and authorization of individual users.
- **Packaging and Delivery:** Prepares the selected collection objects for delivery over a selected channel to a particular customer.

Collection Management defines the business components needed to define and manage the collections of the KB. Individual collections are managed according to defined collection policies. Different categories of metadata are used to manage the collections and support access, e.g. bibliographic, archival and technical metadata. The different Collection Management business components are described below in more detail:

- **Collection Strategy:** Decides which collections to build up and defines the value of the different collections for the KB.
- **Collection Policy Management:** Defines the rules and guidelines for submissions of assets into a particular collection.
- **Metadata Management:** Identifies the different categories of metadata to be associated with particular collections and builds ontologies over the different metadata specification approaches being applied.
- **Delivery and Capture:** Pre-processes digital assets to be ingested: receives or captures digital assets and stores them in a working space for verification in conformance with the defined collection policies.
- **Ingest:** Checks the collection asset for compliance and completeness followed by the archiving of the asset.
- **Collection Storage:** Stores collection assets within the library. In the case of digital material, an Archival Information Package (AIP) has to be maintained in

one or more storage environments as identified in the Reference Model for an Open Archival Information System (OAIS) (ISO 14721:2003).

- **Collection Research:** Extends knowledge and best practices for the development of collections and associated services, including access and presentation.

Preservation Management includes all the business components involved in long-term preservation of digital collections. This is still an active area of research internationally. Two sets of activities are of major importance within this competency. First, one needs to monitor the impact of technology changes over time on the different collections managed, as part of the Preservation Watch function. Second, preservation actions have to be defined to counteract the impact of technology obsolescence, either by migrating collection assets to new formats, emulating obsolete technology or a combination of both. The Preservation Management business components are described below in more detail:

- **Preservation Strategy:** Defines the preservation strategies to be supported by the KB digital repository environment.
- **Preservation Policy Management:** Specifies the preservation policies associated with particular collections or types of digital assets.
- **Preservation Planning:** Defines the actions to implement a specific preservation policy.
- **Preservation Action:** Carries out activities needed to preserve particular collections or types of digital assets: migration (converting collection assets into new formats), emulation (providing new environments to emulate obsolete environments) or a combination of both. Normalisation, i.e. transforming digital assets into formats optimised for the management of particular collections, is also a preservation action.
- **Preservation Research:** Conducts research in the field of digital technologies, network information and the preservation of digital heritage.
- **Technology Monitoring:** Monitors changes in technology environments to be addressed by preservation planning and preservation action.
- **Digitisation:** Specialised preservation action that converts analogue assets into digital assets.

Three business components have been identified as being important to both Collection Management and Preservation Management. These business components are related to the characterisation, validation and cataloguing of collection assets. They are used for initial ingest of a collection but also provide activities which are important for the preservation of the collection. The business components are described below in more detail:

- **Characterisation:** Identifies and records the important characteristics of a digital object and facilitates searching across the characterisation metadata.

- **Validation:** Checks whether the collection assets conform to the associated collection and preservation policies e.g. conformance to file format specifications.
- **Cataloguing:** Builds and maintains metadata to facilitate both general and domain-specific access and searching within or across collections.

Applying the Digital Repository CBM Map

There are a number of ways that the Digital Repository CBM map can be applied within an organisation. The development of the CMB map was triggered by the objectives of KB/IBM Research Group to look into the requirements for an open and integrated preservation framework for the KB to extend the e-Depot, based on IBM’s DIAS solution. The Research Group was aware that any durable electronic deposit solution can never be dependent upon a single vendor providing a closed solution. New formats and preservation tools will continue to be introduced over time, requiring any given solution to be sufficiently open to incorporate functionality from 3rd parties.

The next section will show how the Digital Repository CBM map has been used to identify the generic services to support future developments of the e-Depot. We will then explain how the Digital Repository CBM map can be used to facilitate impact analysis with regard to IT and organisational support.

Identifying Services

The CMB map provides the top-down starting point for the identification of the services that need to be provided by individual business components. Each business component has its own business activities, which can be performed manually or with the support of IT services. The business components drive the definition of potential service candidates. These service candidates are tested for functional usability by determining how they can be used in the different business processes. In order to ensure that a candidate service is reusable across various contexts, it is important to validate it against many different business processes.

This approach will also identify potential “white spots,” i.e. business components not yet supported by any services. White spots are not always a problem. Some business components (e.g. Collection Strategy) might not be implemented through IT services but are, instead, based on human processes resulting in vision documents and associated implementation plans. The strength of the Digital Repository CBM map is the ability to condense major aspects of the library environment into a simple overview that is easy to communicate.

The above top-down approach does not take into account potential existing IT solutions that could provide some of

the required services. The KB has already invested a large amount of money and effort in their current e-Depot solution. Therefore, we also need to evaluate how current IT solutions can provide some of the required services, i.e. bottom-up approach.

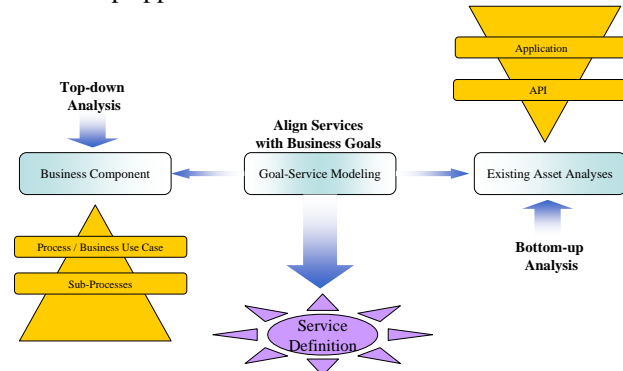


Figure 3: Goal Service Modelling (GSM)

“Goals Service Modelling” (GSM), combines the top-down and bottom-up approaches. Services are identified by the CBM Maps as well as by looking at the functionality of the existing IT solutions. The end result of this exercise is the definition of a complete set of services needed by the organisation to reach the digital preservation goals of its business components, divided in a set of existing services and needed services

Example. One example is the process called *Prepare Content Package for Ingest*. For our exercise we designed this process as in the figure 4. A similar process currently exists in a basic form and creates the Submission Information Package (SIP) to be ingested by DIAS.

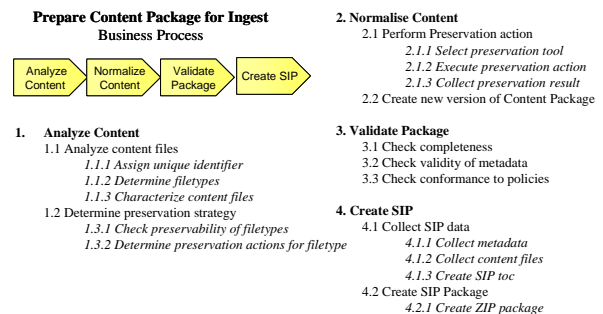


Figure 4: Services - Prepare Content Package for Ingest

Prepare Content Package for Ingest process assumes that Content Packages have been extracted from the Producer Submission Package (PSP). This package is mostly optimized to submit batches, rather than individual digital objects. For example, scientific articles are delivered in batches containing multiple articles of an issue of a periodical, but each article is then ingested individually. As a first step, the content of the packages is analyzed. Each content file, as found in the SIP, will be given a

unique identifier to trace it through the following processes. For each file, an initial determination of file type is generated, and based on this information and the associated policies for the collection a decision is made whether this file needs further characterisation. A conceptual overview of the individual activities is given in figure 4.

After analysing the content file and based on the file type information and possibly other details about the file, as gathered during characterisation, a preservation strategy is determined. First of all, the “preservability” of the file types is determined (whether they are in known and accepted formats, and known risks associated with attempting to preserve given formats), and based on this preservability check, it is determined what preservation actions are needed. An example of such action could be to normalise the file into a preferred file format, such as PDF/A (ISO 19005-1:2005).

The services needed in the business processes will all be attributed to particular business components in the Digital Repository CBM map. For instance, *Determine Filetypes* is a service of the business component Characterisation, and *Collect Metadata* is a service of the Metadata Management business component.

As discussed earlier, there are initiatives across the globe to develop services, which will benefit digital preservation. But how does one determine when a service is viable for an organisation? Which measures are needed to implement services in a manageable way? IBM has developed a Litmus Test to identify viable services, based on the experience of service-oriented architecture (SOA) implementations (where the main goal is to implement services in a flexible and manageable way). The Litmus Test is a set of questions, which need to be answered before implementing a service, from various points of view:

Business Alignment

- Is the organisation willing to fund the service through its lifecycle: provisioning, management, governance, and maintenance?
- Is the organisation willing to share the service internally or externally with clients or business partners?

Composability

- Is the service self-contained (can it be executed within the business components without any resources external to the business component except potential services to be supplied by other business components)?
- Is the service stateless (core operations can be executed as independent transactions)?

Externalized Service Description

- Is the service defined in a way that makes it clear what input and output are expected, and what the

effects of the service will be?

Redundancy Elimination

- Can this service be used by the business stakeholders within all processes where its function is required?

After carrying out the exercises described above – creating a CBM map, identify the services via GSM and performing the Service Litmus Test – an organisation should have an overview of all the services that are needed and where they will be used, as well as which services are generic and reusable. For example in figure 4, the lower level services associated with 1.2 (*Determine Preservation Strategy*), could also be used in other preservation processes, e.g. when reanalyzing already ingested assets that need to be migrated.

Heat Maps

The Digital Repository CBM map can be used not only to identify the services, but also to discuss aspects of organisational and Information Systems (IS) architecture. Recall the definition of a business component as a clustering of business activities with common objectives, which potentially can be managed independently within the organisation.

	Service Management	Collection Management	Preservation Management		
D i r e c t	Release Strategy	Collection Strategy	Preservation Strategy		
	Distribution Plan				
C o n t r o l	Rights Management	Collection Policy Management	Preservation Policy Management		
		Metadata Management ① ②	Preservation Planning		
E x e c u t e	Reading Room ④		Delivery & Capture ③	③ Characterization	Preservation Action ⑥
	Internet	3rd Party	① Ingest	① Validation	Preservation Research
	Licensing and Royalties Management	② Access Management	Collection Storage ①	① Cataloguing	Technology Monitoring
	Packaging & Delivery ①		Collection Research		Digitalisation

① DIAS ② KB Catalogue ③ Electronic Post-Office ④ Reference WorkStation ⑤ Dioscuri ⑥ UVC

Figure 5: Example of IS Heat Map

Heat Maps are a visualization tool to map different types of needed resources onto the identified business components. They illustrate points of potential resource conflict between different business components.

The above example of a Heat Map shows how the major applications of the KB are positioned to support the business components. The IS Heat Map provides an overview of the different applications currently supporting KB’s business objectives. It also highlights potential white spots. For example, technology monitoring at the moment

seems not to be supported by any applications in the operational environment. An operational solution could be implemented in one of the next versions of DIAS based on Preservation Manager Proof of Concepts (Oltmans, Diessen and Wijngaarden 2004), using another file format registry or the implementation of the results in this area of the PLANETS project. Normally the business components would be coloured to represent their state. In the above IS Heat Map the colour would be used to represent the fit between actual and required IT support for the business component, e.g. bad, average and good.

The Heat Map approach can also be used to evaluate whether the current organisational structures are aligned with changing requirements introduced by the management of digital collections inside an organisation. Ideally, responsibility for a given business component will not be divided over multiple organisational units. The responsibility for any business component should, based on the definition, be the responsibility of one organisational unit to maintain clear responsibilities for the resources and actions to be executed by the business component. In practice, such cross-unit sharing of a component often does occur, which can generate additional risks and coordination costs.

The Digital Repository CBM map and Heat Maps can be used together to compare different organisations and digital repositories. As discussed above, several current initiatives are investigating the characteristics of trustworthy repositories, along with criteria for their audit and certification. The use of CBM, Heat maps and GSM could be used to translate these criteria into digital preservation environment solutions in specific organisational and institutional contexts.

Conclusions and Next Steps

Key to the success of any digital repository focused on long-term preservation of its collections is openness and adherence to open standards. Technology innovation is only accelerating, with new digital formats and supporting application software being introduced and digital objects becoming more complex. Digital repositories will need to adapt continuously, in order to support these new formats with appropriate services for characterisation, validation, ingest, preservation and access. At the moment, the preferred industry approach to make systems flexible is to adopt a service-oriented architecture (SOA) and associated web service standards.

The Digital Repository CBM map enables systematic analysis of the impact of new developments. We believe it is important to generalize and refine the Digital Repository CBM and discuss the results within the digital preservation community. The CMB map could also provide added value to ongoing audit and certification initiatives.

We have shown how the top-down Digital Repository CBM map provides the required context by identifying the business components needed to manage a digital repository. With GSM it is possible to identify the services. The Heat Maps show the results of impact analyses that can be facilitated by the CBM map for a multitude of factors: services, IS, resources, processes and organisation. We consider this exercise a useful method to model required services in the future.

The next steps will focus on the communication and validation of the initial Digital Repository CBM map by the long-term digital preservation community. We also want to evaluate the effectiveness of the Digital Repository CBM map in the comparison and coordination of multiple repository environments.

References

- Dobratz, S.; Hänger, A.; Huth, K.; Kaiser, M.; Keitel, C.; Klump, J.; Rödig, P.; Rohde-Enslin, S.; Schoger, A.; Schröder, K.; and Strathmann, S. 2006. *Catalogue of Criteria for Trusted Digital Repositories, Version 1*. Frankfurt, Germany: nestor Working Group on Trusted Repositories Certification.
- Ernest, M.; and Nisavic, J.M. 2007. Adding Value to the IT Organization with the Component Business Model. *IBM Systems Journal* 46(3): 387- 403.
- ISO 14721. 2003. Space Data and Information Transfer Systems – Open Archival Information System – Reference Model.
- ISO 19005-1. 2005. Document Management – Electronic Document File Format for Long-Term Preservation – Part 1: Use of PDF 1.4 (PDF/A-1).
- McHugh, A.; Ruusalepp, R.; Ross, S.; and Hofman, H. 2007. Digital Repository Audit Method Based on Risk Assessment (DRAMBORA). Digital Curation Centre and Digital Preservation Europe.
- Moore, R. 2008. Towards a Theory of Digital Preservation. *International Journal of Digital Curation* 3(1): 63-75.
- Oltmans E.; Diessen R.J. van; and Wijngaarden H. van 2004. Preservation Functionality in a Digital Archive. In Proceedings of the Joint Conference on Digital Libraries, 279-286. Tucson, Arizona: ACM.
- Trustworthy Repositories Audit & Certification (TRAC): Criteria and Checklist*. 2007. Chicago, Ill.: Center for Research Libraries and OCLC Online Computer Library Center.

Development of Organisational and Business Models for the Long-Term Preservation of Digital Objects

Tobias Beinert*, **Susanne Lang***, **Dr. Astrid Schoger***
Prof. Dr. Uwe M. Borghoff**, **Dr. Harald Hagel****, **Michael Minkus****, **Peter Rödiger****

*Bayerische Staatsbibliothek / Bavarian State Library,
Ludwigstraße 16, D-80539 München,
<FN>.<LN>@bsb-muenchen.de

** Universität der Bundeswehr München / University of
the Federal Armed Forces Munich, Werner-Heisenberg-
Weg 39, D-85579 Neubiberg, <FN>.<LN>@unibw.de

Abstract

The number of digital objects (and digital collections) will increase rapidly within the next years since mass digitisation activities have started all over the world. Although it is obvious that these objects are of enormous scientific and cultural value, some crucial aspects of ensuring their long-term preservation and access to them have so far not been thoroughly addressed. This means that there is an urgent need for developing (and implementing) new and reliable models in order to deliver a sound organisational and financial framework for institutions (and enterprises), that are concerned with digitisation and long-term preservation of digital objects. To this purpose the Bavarian State Library (BSB) and the University of the Federal Armed Forces Munich, are carrying out a study, funded by the German Research Foundation (DFG), that explicitly addresses the perceived shortcomings by analysing the current state of long-term preservation in Germany, developing solutions in the form of scalable business and organisational models and clarifying the agenda for further research.

Background/Motivation

Today, the access to digital cultural heritage is a granted service of the traditional memory organisations. No longer only small projects on the digitisation of specific scientific and aesthetic values of the stocks of our organisations are realized. The focus is rapidly shifting away from pure boutique to mass digitisation projects with several thousands of titles. To secure the availability of this content for the long term is one of the priority tasks of memory organisations.

Long-term preservation of the underlying data has been recognized as an absolute necessity, yet infrastructures can change, funds run dry. Therefore sustainable structures have to be created to ensure the preservation of our digital heritage in every case. Apart from reusable technical solutions, in particular stable organisational, legal and financial models have to be developed, which can be harmonised in a strategy for long-term preservation of digital content.

Objectives

For that reason, the Bavarian State Library (BSB) and the University of the Federal Armed Forces, Munich are

carrying out the study 'Development of and Organisational and Business Models for the Long-term Preservation of Digital Objects from DFG (German Research Foundation)-funded digitisation projects'. The intention is to demonstrate how organisational and Business Models can be designed and realised for the long-term preservation of digitised material and where further research and development has to be done.

Comparing the general aims and the current state of the art of long-term preservation (LTP) shall provide the needed guidelines for an in-depth investigation into the four dimensions organisation, technology, finance, and law.

Therefore the first sub-goal of the study is a detailed description of the actual situation in digitising and archiving institutions in Germany. Besides the analysis of relevant reports and studies a purpose tailored questionnaire serves as a basis for deepened research in the named dimensions.

Concerning the dimension organisation we are planning to present possible Organisational Models for long-term preservation. A methodological framework in form of a Process Model is going to be designed first in order to create the basis for the development and evaluation of Organisational Models.

In a next step existing technical solutions for long-term digital archiving and their advantages and disadvantages are presented and assessed. In the area of finance the possibilities of income generation should be explored and potential savings identified. A corresponding examination of the legal framework for innovative business and Organisational Models is also part of the study.

In a final step the need for further action in the dimensions of organisation, technology, finance and law will be pointed out. The developed models will thereby provide the opportunity to clearly pinpoint and define the problems of long-term preservation. Finally, a roadmap for planning studies and projects can be drawn more precisely.

The scheduled timeframe is from January 2008 to January 2009. This article gives an overview of first results of the study and further expected outcomes.

Approach

Development and assessment of organisational and Business Models require a solid methodological foundation since the long-term preservation of digital information is a quite complex task. Different business goals as well as technical, legal and financial opportunities and constraints lead to numerous possible system configurations with many interdependencies. Adequate and clearly defined models will assist to describe, analyse, and design complex technical and organisational systems. Fortunately, we can build on models and frameworks already applied or under development in economics, administration, or even in long-term preservation.

Methodological Approach at a Glance

First, we adapt basic definitions of published Business Models in order to get a generic Business Model

appropriate as starting point for our study. Business Models mainly provide methodological support for achieving business goals. They also consider the context of a business like the situation for market and for competition. As memory organisations generally deal with public goods, we have to bear in mind that their situation is extensively shaped by national and international legislation. The legal dimension will be investigated by a corresponding expertise.

Then we present a procedure to get a generic Process Model which enables us to describe the numerous results of prior conceptual work in a consistent and structured way. For example, several models for digital libraries as well as for long-term preservation are already published, but in general they use their own languages and focus on different aspects. Of course, we also consider prior work that is not specific to long-term preservation or digital libraries like generic models for activity based accounting or information lifecycle. The standardised description helps us to find out gaps, inconsistencies, and useful results of prior work relevant for developing Business and Organisational Models. Moreover, the Process Model provides an additional schema for describing the elements of a specific Business Model precisely.

In order to develop Business Models for memory organisations we need another set of procedures that assist to introduce aspects specific to long-term preservation and to map the current practices of the numerous DFG-funded digitisation projects to the elements of a Business Model. Therefore, we refine the generic Business Model. In order to get a realistic picture of current practices we designed a questionnaire which

reflects the elements of process and Business Models. The questionnaire should also provide future visions for the distribution of information.

So we think to obtain an adequate set of tools as well as enough information from practice to develop and to assess Organisational Models. Thus, we will be able to take into account individual business goals as well as technical, financial, and legal dimensions. The Organisational Models will show how static and dynamic structures can be designed and implemented.

Approach in Detail

Procedures for Deriving Generic Business Models

Of course, we initially need a basic understanding about Business Models. We found definitions in literature that seem to be adaptable for our purposes. Timmers¹ defines a Business Model as architecture of product-, service- and information-related business processes. It comprises a description of the participants, their roles and their

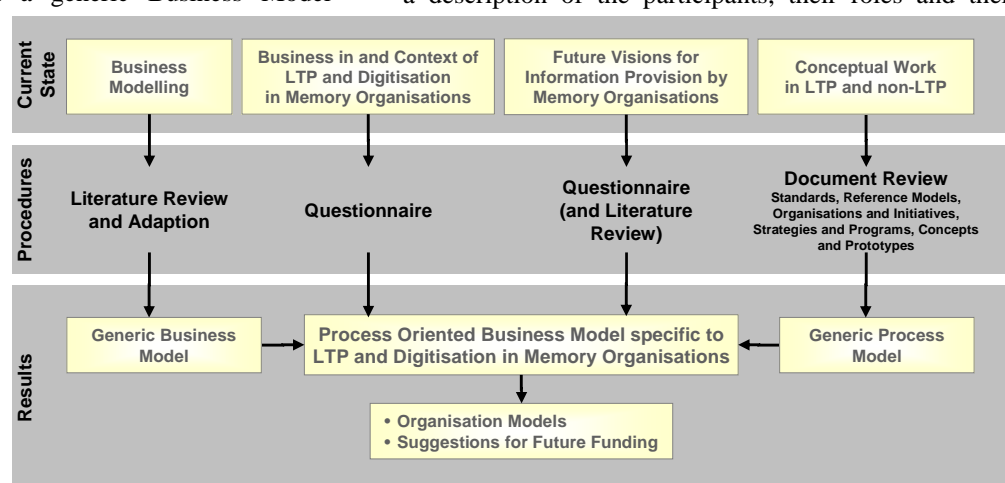


Figure 1: Methodological Approach

potential of benefit as well as a description of sources of proceeds. According to the definition of Porter² a competitive strategy is the precondition for a Business Model. Here, competition is considered as the continuous process of seeking new and better ways of satisfying needs in order to increase one's own prosperity. The competition strategy and the business goals are prerequisite for a Business Model whereby the situation of competition has to be considered. Finally, the business plan specifies how the business goals should be realised. Figure 2 gives an overview of our adaption called generic Business Model.

Procedures for Deriving Generic Process Models

Process modelling is an additional method for controlling the complexity of long-term preservation. Process modelling is a well established method for designing and reengineering complex systems in administration and industries. In order to get input for our models we are analysing current conceptual work from the LTP community as well as from other areas with assumed relevance for long-term preservation. In order to analyse

¹ Timmers 1999, p. 23-27.

² Porter 1996, p. 23-28.

all the documents and their content (a first inventory has revealed some thousand documents) and to facilitate the mapping onto the elements of a Process Model and finally of a Business Model we have introduced the following categories:

- Reference Models are relatively abstract and general models, which are also characterised as conceptual frameworks. They form a basis for a common understanding and for specific models, e.g. OAIS³.
- Standards represent the state of the art built on the principles of fairness, consensus, and documentation. They simplify the comparability, assessment, and interoperability of products, systems, and services, e.g. XML, ISO 9000.
- Organisations and initiatives reflect domain specific as well as integrative aspects having the big picture (missions) in mind. They develop strategies and cooperations, e.g. DFG⁴, Library of Congress⁵, Nestor⁶.
- Strategies and programs form the frame for concrete activities or projects, e.g. the National Digital Information Infrastructure and Preservation Program (NDIIPP)⁷.
- Basic projects develop concepts and prototypes, evaluate concepts and practices, and conduct research, e.g. LIFE⁸, TRAC⁹.

The next step deals with the mapping of the conceptual work onto the elements used in Process modelling. These elements cover tasks, task performers, resources as well as static and dynamic structures (the organisational structure and the procedural organisation), managed information, and finally spatio-temporal and quantitative aspects. Therefore, Process Models provide a consistent description of the concepts in memory organisations. We

also consider concepts that are not directly related to long-term preservation like information life cycle models

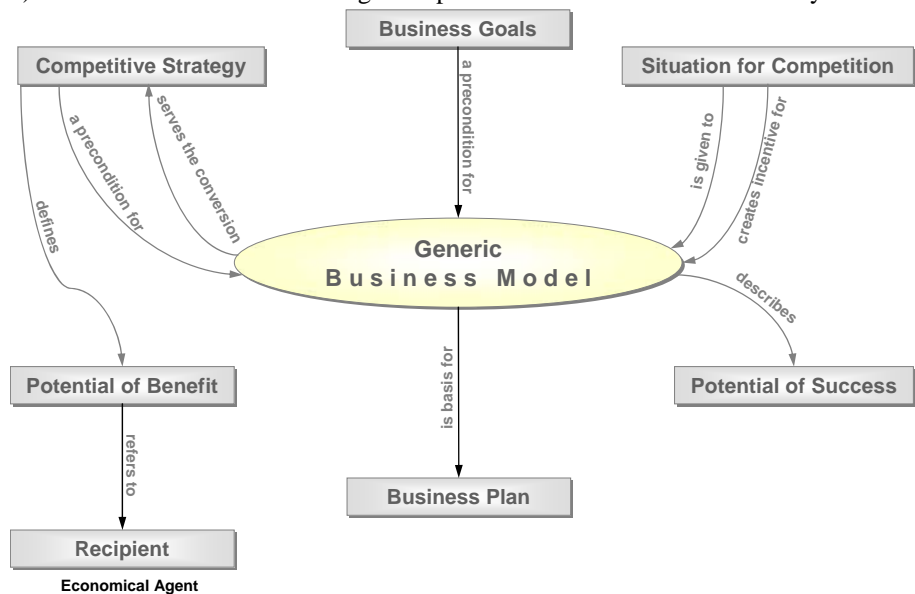


FIGURE 2: Generic Business Model

or service oriented architectures.

So we will get a structured description that allows us to recognise interdependences, gaps, and inconsistencies. Especially, we are interested in parts that can be reused for Process modelling. For example, the functional entities and sub-functions as specified in the OAIS reference model can be transferred into a generic Process Model. According to common practices we introduce three different types of processes. First, management processes direct and control all the other processes¹⁰. Second, core processes realise the goals of an

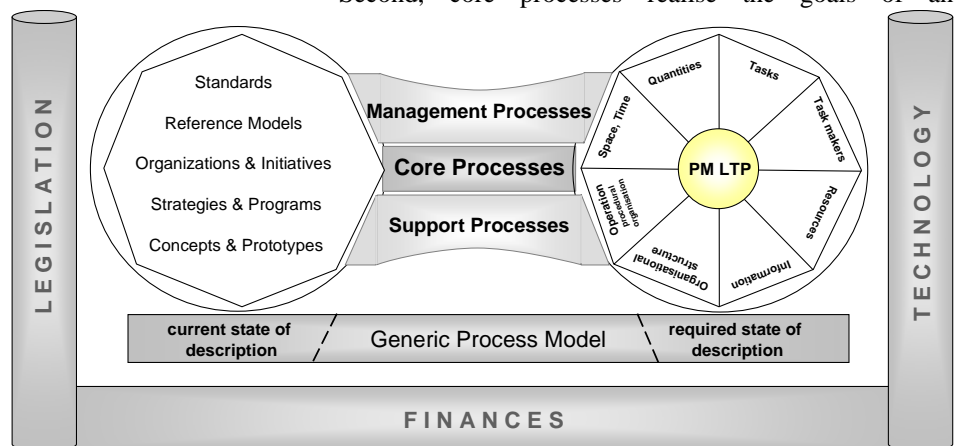


Figure 3: Procedures for deriving generic process models for LTP (PM LTP)

organisation. These processes are also value adding from a customer's point of view. Ambitious business goals usually require interdisciplinary core processes. Third, support processes do not directly add value for the customer, but they are necessary for the core processes to work properly. These types of processes are primarily elements for outsourcing.

The generic Process Model enables us to make substantial statements concerning the dimensions,

¹⁰ Management processes are directly addressed by ISO 9001.

³ CCSDS 2002.

⁴ <http://www.dfg.de/en/index.html>.

⁵ <http://www.loc.gov/index.html>.

⁶ <http://www.langzeitarchivierung.de>

⁷ <http://www.digitalpreservation.gov>.

⁸ <http://www.life.ac.uk>.

⁹ CRL 2007.

technology, finances, and law. But the model is still largely independent of concrete Business Models. Figure 3 illustrates the procedure for deriving the generic Process Model.

Procedures for Deriving Process Oriented Business Models Specific to LTP and Digitisation

In order to move from the generic models to models specific to the digitisation and long-term preservation of digital objects further procedures are required. First, we need a clear picture of the current practices and existing visions, and second, we have to adapt and populate the generic models.

Questionnaire

The questionnaire is an important milestone in our approach to get a realistic picture of how memory organisations currently manage and operate their digitisation projects and long-term preservation. The design of the questionnaire mainly considers aspects relevant for business and Organisational Models. But for reasons of acceptance we reduced the explicit use of technical terms. Especially, we assume that publicly funded memory organisations do not think primarily in abstract terms common in Business modelling. Anyhow, we are convinced that good practices and even future visions concerning these subjects already exist.

Recipients of the Questionnaire

Whereas the analysis of existing conceptual work was highlighting and outlining mainly theoretical aspects of long-term preservation, the questionnaire is supposed to give us a more practical view on the current situation in digitising and long-term archiving institutions. As a result tangible input for the development of organisational and Business Models can be provided.

By selecting the projects to be surveyed an intentional effort has been made to cover the widest possible range of digitisation and preservation projects. So the varying characteristics of internal organisation, process cycles and workflows can be detected more clearly. Although initially the focus of the survey had been limited to single digitisation projects, it soon became clear that the scope had to be extended to the institutions as a whole as stable organisational structures can only be identified and evaluated in an overall context.

Selection criteria were inter alia the nature of the institution responsible for the project (library, archive, museum, research institution) and its experience in the area of digitisation and long-term preservation. Furthermore it was important, when and for how long a project was realised, how many digital objects were produced, what was the original material for the digital media and how it is made accessible for the designated user community.

Since the study aims to develop widely usable business and Organisational Models, the survey did not only address the well known players of long-term preservation in Germany but also specific small and medium-sized organisations with limited budgets and lower levels of experience in this field. It was also relevant whether the projects were conducted independently or in cooperation with other institutional partners and private enterprises.

The coverage of the institutions surveyed ranges from the highly specialized digitisation centres in Munich and Goettingen to medium-sized institutions such as research institutions with special interests in digitisation to smaller foundations which have until now just converted and archived parts of their photo stocks into digital forms.

Conception of the Questionnaire

The questionnaire has been split into a general area 'institution' in which general information about the digitisation and long-term preservation process in the various institutions has been queried and a more specific part with particular questions to the individual projects. Initially we asked basic questions about the institutions' general motivation for digitisation and long-term preservation, selection criteria for the material used, responsible departments and persons and the number of already digitised and archived objects. Issues of interest were also the orientation or adherence of special guidelines and the development of institutional digitisation or long-term preservation policies.¹¹

The answers to these questions should give us a basic overview of the surveyed institutions and their experiences in this field. The aspects relevant for the development of our models like fields of activity, specific tasks, tasks managers, financial means, human resources, used material, internal process structures and workflows, time, space and quantity have been queried according to the dimensions organisation, personnel, technology, and finance in the course of the questionnaire.

Apart from these fundamental issues, we have put special emphasis on the subject of customer orientation of digitisation organisations by asking for offered services and products. In this part of the survey we were also interested in possibilities of exploitation of digital objects, generation of revenues and ways of refinancing digitisation and long-term preservation.

In the final part of the questionnaire, 'visions', the respondents were given the possibility to outline future prospects for their institutions and specify further general needs concerning the fields of R&D, cooperation and consultation services in long-term preservation.

Procedure for Deriving a Business Model for LTP

Now we are going to explain why and how the generic Business Model is specialised for long-term preservation. Of course, this may be a first step of iteration because the results of the questionnaire are not yet fully available. Let us start with elements for a LTP-specific Business Model that are already on-hand.

Fundamental Business Goals of Long-Term Preservation

The fundamental goals of long-term preservation have been articulated officially several times and are known among experts.¹² More particular targets derive of course from the general mission and legal obligations of the

¹¹ IFLA 2002, DFG 2008 et al.

¹² UNESCO 2003, Nestor 2006 et al.

individual institutions. In order to identify the required fundamental business goals it is more useful to draw on the concepts of Trusted Digital Repositories, which specify the goals of long-term preservation much more precisely.¹³ According to these concepts the overriding and action-guiding principle for digital repositories is to secure integrity, authenticity and availability of digital objects. To evaluate if and how a digital repository is able to fulfil this main task, its effectiveness and efficiency have to be analysed and in a second step optimised. While effectiveness deals with the question whether a repository can preserve digital objects for the long-term at all, efficiency rates the benefit-cost-relation of the used resources. Beyond these main business goals also other targets, such as the creation of transparent decision-making criteria for selecting the material to be digitised have to be taken into account, if strategic models for long-term preservation are to be developed. Sustainable organisation and cooperation structures have to be created and consolidated to enable memory organisations to carry out their duties in the field of long-term preservation in an effective and efficient way. Of course, fundamental business goals are useful for

Wirtz¹⁴ to facilitate the design of cooperative organisation forms. We separate the original Partial Model named 'Production and Procurement Model' in two models. We also split the original Partial Model named 'Organisation Model' in two models named Utilisation Model and Operator Model and add some extensions. In summary we have: Market Model, Product Offer Model, Production Model, Distribution Model, Utilisation Model, Procurement Model, Operating Model and Capital Model. Each of these Partial Models is described by the elements of the generic Process Model. Now we have the granularity necessary for scaling and tailoring systems according to needs of individual memory organisations.

The granularity allows analysts to isolate and consider specific aspects without losing track of all the interdependencies. Starting points for optimisations or innovations as well as externally induced changes can be identified and systematically assessed. For example, starting with tasks, processes including related actors can be identified, and required resources can be determined. Ideally, key figures allow analysts a quantifiable assessment of different configurations.

Of course, this study cannot provide off-the-shelf Organisation Models. There are too many different individual situations and too many possible configurations. But the models, that we are developing, will enable institutions to design and to assess their business and to formulate business plans. The models will also help to evaluate practices and to design patterns that can easily be reused.

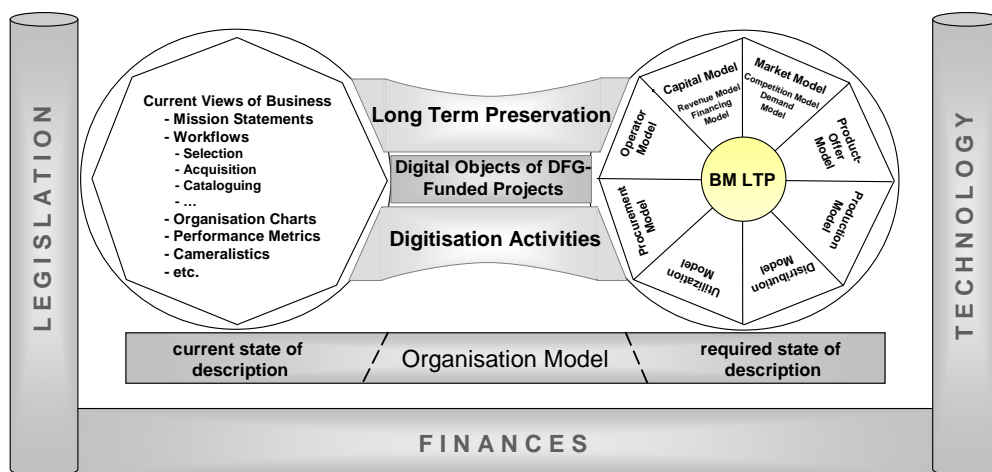


Figure 4: Procedures for deriving Business Models for LTP (BM LTP)

developing Business Models, but they cannot provide enough information for economical analysis. Moreover, the current state of describing business in LTP and digitisation is not suited for economical analysis in general. But often existing documentations of the running business comprise information that can be made explicit for economic analysis by applying adequate Business Models as depicted in figure 4. Finally, having the information in form of a process oriented Business Model allows individual institutions to assess if market needs are satisfied effectively and efficiently. This Business Model will then be combined with the Process Model. The next sections show more details.

The Process Oriented Business Model in Detail

In addition to the presented process oriented view (see figure 3) we adapt the generic Business Model (see figure 2) in order to cover the complexity of long-term preservation and to derive Organisation Models. Therefore, we adapted the Business Model published by

(1) The **Market Model** consists of the Demand and the Competition Model. The Demand Model identifies consumers and classifies them into market segments according to the Utilisation Model. The segmentation helps to optimise the offer. The Competition Model identifies for every sales market its competitors and its roles and relations. Profit centres represent the consumers on internal markets, end customer are the external demanders. The Market Model serves to determine for every sales market one's own opportunities on the market.

We have to take into account that the market for memory organisations is extensively regulated by the legal framework on the one hand and the fact that the awareness of existing markets is obviously not very high on the other hand.

(2) The **Product Offer Model** determines the service portfolio that is adapted to the individual needs of the actors. How actors can use the products and how the

¹³ CRL 2007, Nestor 2006.

¹⁴ Wirtz 2000.

utilisation is technically and non-technically supported characterise the offered services. The Product Offer Model is closely linked to the Utilisation Model and shows the specific use by customers within different market segments. The required sustainability of the offered services in LTP is distinctive to usual product offers. The quality of service is endangered by technical obsolescence and loss of context.

(3) The **Production Model** describes the stepwise transformation of products or services adding value. For every transformation the input and output has to be specified. Additionally, all resources required by each transformation are listed in this model. Outputs have to fulfil the specified quality standards, which consider views from different Partial Models (e.g. costs or product properties).

In our case all transformation processes that maintain integrity, authenticity, and accessibility of information are also part of this model. Of course the enhancement of information provision, for example by adding descriptive metadata to the digitised objects, is also adding values. In comparison to usual products the specification of information's quality is hard to formulate.

(4) The **Distribution Model** specifies how the offered products or services get to the customers, and it informs about costs, time, and quality of the distribution process. The distribution includes internal as well as external consumers. The model distinguishes the products into two groups: material and immaterial goods. Even in the digital world we have to consider the handling of extreme valuable masters.

Ubiquitous computing and growing bandwidth for communication will lead to new Distribution Models.

(5) The **Utilisation Model** serves for the identification and description of the actors in form of a role-model, with which the relations between actors and elements of the Process Model can be declared.

The model focuses on the internal as well as on the external users' view and bridges the gap between users' wishes and the products and services that can be offered economically reasonable. The model should help to recognise the willingness to pay for products and services and therefore provides input for the Revenue Model.

We assume that it will be hard for memory organisation to estimate the willingness to pay for public goods that were guaranteed by law to be free yet.

(6) The **Procurement Model** identifies and describes the raw materials (external inputs) and the factors of production necessary to run the transformations as described in the Production Model. In general, the procurement is also subject to market mechanisms.

It also includes masters, licences, and IT-components. Procurement and running of mission critical IT-infrastructure are not the original core business of memory organisations. Therefore they need the capability to specify their requirements and to manage and control procurement of products or services.

(7) The **Operating Model** in the broad sense describes alternatives of production as well as product offering by third parties. The model comprises the internal and the external relationship between partners.

Leveraging the specific competence of an external partner can lead to more efficiency and effectiveness. Partners can share innovations and resources as well as risks. Organisations can concentrate on core processes and can draw off resources as actually required by business. Even small units can benefit from partnerships, because they need not to operate a complete infrastructure.

Operating models are characterised by two types of relationships. Internal relationships are established to run a business as a whole, while external ones focus to the world outside this cooperation. Achieving a win-win-situation is usually the driving force for establishing cooperation. The types of cooperation may range from owner-operated units to the outsourcing of complete core processes.

All forms of partnerships require thorough consideration, especially if mission critical subjects are involved in the long term.

(8) The **Capital Model** consists of the Financing and the Revenue Model. The Capital Model requires a description how to manage and control the inflow and outflow of resources. It completely lists sources of revenue and facts that cause expenditure. Therefore the model has also to show all operational areas that are indispensable on the one hand and that cannot realise revenue on the other hand.

Such a level of transparency is the basis for a concrete business plan and the prerequisite for any kind of reengineering (cost-benefit analysis).

Some models focus on expenditures (e.g. Procurement, Production, Product Offer, and Distribution) whereas others focus on revenues.

With the help of the eight Partial Models innovative future business ideas in the field of information and communication can be identified in a method- and model-driven way. This multi-perspective view puts us in the position to clearly assign the constitutive elements within the framework of the Organisational and Business Model and to adapt them to different conditions by brainstorming.

The mentioned models have to be described in detail to enable substantial statements on the subjects of technology, finances, and law - either for the particular model or for the systematic combination of several models. In order to transfer the rather abstract model to an individual memory organisation the model has to be instantiated.

We call this instantiated form 'Organisation Model'. Now the actual tasks of existing institutions can be assigned to real task performers and interdependences between them and others elements of the Process Model, especially resources, can be recognised.

An Example - from Workflows to the Business Model at the BSB Munich

The following paragraph exemplifies two of the Partial Models by taking a deeper look at some aspects of the organisation of digitisation and long-term preservation at the Bavarian State Library (BSB).

Since the foundation of the Munich Digitisation Centre in 1997 the Bavarian State Library became one of the major content providers among libraries in German-speaking countries, now hosting already more than 30,000 volumes and approximately 10 Million pages. There will be an enormous growth of the collection within the next few years as several important new digitisation activities have been started in 2007. These digitisation activities include amongst others, the so-called VD-16-digital-project, in which automated scanning technology for the digitisation of books of the 16th century is applied and the public private partnership with Google. More than one million books out of the copyright-free holdings of the library are going to be scanned by Google and will be hosted by Google Books as well as in the Digital Collections of the Bavarian State Library for free access.

The Production Model at the Munich Digitisation Centre includes four main steps: a) Image Capture b) Indexing and Access c) Publication d) Storage and Preservation. Due to the limited context of this article only the part of Storage and Preservation can be described in detail.

After the first steps of the digitisation-workflow have been successfully completed, the digital master images and the corresponding metadata are being transferred to the Leibniz Supercomputing Centre (LRZ).¹⁵ Its powerful technical infrastructure is being used for the archival storage of the Bavarian State Library and thus delivers a major contribution to the preservation of the added value that had been created by the transformation of an analogue into a digital object. This partnership of BSB and LRZ contains also basic elements of the Operating Model and has to be analysed from that point of view in a further step of our study.

The Leibniz Supercomputing Centre uses a TSM/HSM storage system based on a tape library for archiving BSB's digital content. Incoming digital objects are automatically stored in this archival system every day.

Each digitised volume is been kept as an uncompressed master copy together with the complete bibliographical metadata and basic technical metadata information. Put together this makes a 'self-explaining' archival information package which remains usable even in case of loss of all external reference systems (e.g. the database or the local catalogue system). For the efficient storage of large amounts of data Hierarchical Storage Management (HSM) is being used. Several storage systems with different quality of services are integrated into a single file system view. According to defined rules the files are automatically and transparently migrated between the storage layers. Virtually there is no limit for the amount of stored data as the HSM file systems usually use magnetic tapes as final storage layer. However, especially when HSM is used for long-term

archiving with a quickly and continually increasing number of files, the performance of meta-data operations (e.g. identifying files for migration) could become a critical issue. This means, that in order to keep the whole system manageable additional measures (e.g. survey of file formats and file numbers) have to be taken. The very efficient architecture of the archival system makes it possible to locate and retrieve every stored file in just about two minutes time. Only widespread and well documented file formats (TIFF, JPG, PDF/A and plain text files, e. g. XML) are being stored in the archive. For that reason, special preservation activities (e. g. format migration, emulation) have so far not been necessary, but can easily be implemented if needed. A first hardware migration of the complete data stock of the library was completed successfully by January 2007 (then 42 TB).

The Distribution Model describes the ways and modalities in which digital objects and associated services are brought to the user community. The basic means of delivering digital objects to the users of BSB is obviously the WWW. Several options of accessing the digital content are offered by the Bavarian State Library. The user can either use search engines on the WWW (e.g. Google), global or regional catalogues (e.g. WorldCat, Gateway Bayern) or he can search the library's local catalogue system (OPAC) where a link inside the bibliographic record will lead him directly to the digital object. Another way is to browse or search inside the digital collection's homepage.¹⁶ There is a special server-infrastructure which processes user requests, so that it is not necessary to revert to the archived objects at the Leibniz Supercomputing Centre. Although the vast majority of digitised books is freely accessible on the web, in some cases access has to be limited to in-house-usage at the public Internet PCs of the library's reading rooms due to copyright restrictions. The basic access format is JPG, but the user can also generate a PDF version for a greater flexibility in handling and printing the objects.¹⁷ Every image is available in two or three sizes for different zoom levels depending on the size of the book.

Besides online distribution of already digitised material the BSB also provides the opportunity of a 'Digitisation-on-Demand'-Service, which enables the user to order a digital copy of almost every printed book out of the library's depository. In this case the user can chose his preferred form of delivery: paper copies, CD / DVD or Internet Download. Additionally high resolution images can be ordered for special scientific or commercial purposes. Depending on quality, size, colour, quantity, processing of special requests, intended use and form of delivery fixed fees are charged. This basic pricing model can be a good first starting point for the further development of more elaborated Financing and Revenue Models for the BSB in particular as well as for memory organisations in general.

¹⁵ <http://www.lrz-muenchen.de>

¹⁶ <http://www.digital-collections.de>.

¹⁷ This service is in trial operation and is being introduced step by step for all digitisation projects funded by the German Research Foundation.

Conclusion

What we can see from the first questionnaires that have been returned so far, is that all our interviewees, asked for their future needs, first and foremost would like to have a source of specific advice on questions arising from practice as well as generally accepted standards in all the dimensions mentioned above - organisation, technology, finance, and law. That basically means feasible workflows tailored for day-to-day-business, suggestions for the adoption and application of metadata and technical standards and help with copyright issues. Furthermore financial support for the establishment of sustainable human resources structures for long term preservation is considered to be of special importance.

On the basis of the methodical approach – from the Generic Business Model and Generic Process Model to the Process Oriented Business Model for digitisation and LTP in memory organisations and through instantiation to a specific Organisation Model for a particular memory organisation - we are able to get a holistic view on digitisation and LTP. Conceptual work and aspects of real memory organisations can be systematically analysed. New models for different contexts can be derived, and existing models can be optimised methodically. The process oriented approach facilitates a systematic reengineering.

Our approach is flexible enough that it can also be applied to other types of digital objects, like scientific data or multimedia contents, long-term preservation outside memory organisations and in an international context.

References

- CCSDS- Consultative Committee for Space Data Systems (2002): *Reference Model for an Open Archival Information System (OAIS)*. <http://public.ccsds.org/publications/archive/650x0b1.pdf> (2008-08-14).
- CRL, The Center for Research Libraries and OCLC Online Computer Library Center, Inc., eds. (2007): *Trustworthy Repositories Audit & Certification: Criteria and Checklist*. <http://www.crl.edu/PDF/trac.pdf> (2008-08-13).
- DFG (German Research Foundation) (2008): *Praxisregeln im Förderprogramm 'Kulturelle Überlieferung'*. http://www.dfg.de/forschungsfoerderung/wissenschaftliche_infrastruktur/lis/aktuelles/download/praxisregeln_kulturelle_ueberlieferung_0208.pdf (2008-08-13).
- DIN EN ISO 9001:2000 (2000): *Qualitätsmanagementsysteme – Anforderungen*. Berlin: Beuth
- IFLA (2002): *Guidelines for Digitization Projects*. <http://www.ifla.org/VII/s19/pubs/digit-guide.pdf> (2008-08-14).
- Nestor - Network of expertise in long-term storage (2004): *Digitale Langzeitarchivierung und Recht*. http://www.langzeitarchivierung.de/downloads/mat/nestor_mat_01.pdf (2008-08-13).
- Nestor - Network of expertise in long-term storage (2006): *Memorandum on the long-term accessibility of digital information in Germany*. <http://www.langzeitarchivierung.de/downloads/memo2006-e.pdf> (2008-08-13).
- Nestor - Network of expertise in long-term storage (2006): *Catalogue of Criteria for Trusted Digital Repositories, Version 1 (draft for public comment)*. <http://edoc.hu-berlin.de/series/nestor-materialien/8/PDF/8.pdf> (2008-08-13).
- Porter, Michael E. (1996): *Wettbewerbsvorteile. Spitzenleistungen erreichen und behaupten*. Frankfurt: Campus-Verlag.
- Timmers, Paul (1999): *Electronic commerce. Strategies and models for business-to-business trading*. Chichester et al.: Wiley.
- UNESCO - United Nations Educational, Scientific and Cultural Organization (2003): *Charter on the Preservation of the digital Heritage*. http://portal.unesco.org/ci/en/files/13367/10700115911Charter_en.pdf/Charter_en.pdf (2008-08-13).
- Wirtz, Bernd W. (2000): *Electronic business*. Wiesbaden: Gabler.

Long-term Preservation of Electronic Literature

Sabine Schrimpf

Deutsche Nationalbibliothek
Adickesallee 1
60322 Frankfurt
s.schrimpf@d-nb.de

Abstract

Authors have always been looking for new and innovative ways of aesthetical expression and they have been quick to take advantage of the new possibilities, offered by the World Wide Web and the Internet. Today, there is a respectable community – in terms of size and quality – of “digital poets”, who publish their texts in the internet and whose literature above all shares one mutual feature: a prominent and crucial use of computer technologies.

Being the cultural heritage for future generations, electronic literature is worth preserving, just as any other form of contemporary literature. Yet, due to its use of interactive and dynamic elements and reliance on the latest technology, contemporary electronic literature is extremely vulnerable and difficult to document. While archives, libraries and museums are still trying to develop preservation strategies for electronic literature, many of the early works have already volatilized. In Germany, two institutions and a cooperative network have joined forces in order to address this challenge concertedly: nestor, the network of expertise in long-term storage of digital resources, the German Literature Archive, and the German National Library.

Background

As opposed to the preservation of research data, no comprehensive, international efforts address the preservation of electronic literature. So far, there have only been a few initiatives in the USA which address the need to preserve electronic literature, such as the Preservation, Archiving, and Dissemination (PAD) project of the Electronic Literature Organization or Archive-it, a collaboration of the Library of Congress, the Electronic Literature Organization and the Internet Archive. These projects are still in their early stages of development.

In Germany, the Deutsche Literaturarchiv Marbach (DLA, German Literature Archive Marbach) is responsible for collecting, archiving and making available contemporary German literature. Primary sources and secondary literature are collected as comprehensively as possible. Since 1997, the DLA has expanded its efforts to electronic literature, beginning with the inclusion of the collecting field “e-journals”.

The Deutsche Nationalbibliothek (DNB, German National Library) is the legal deposit library for all German and German-language publications since 1913.

As of 2006, native digital publications are also included in the German legal deposit law.

In cooperation with nestor, the German network of expertise in long-term storage of digital resources, DLA and DNB have taken the initiative to develop a preservation strategy for electronic literature in Germany. In March 2008, they managed to bring together the relevant stakeholders in the German National Library in Frankfurt: authors, archivists, librarians, and legal experts met to discuss the challenges of long-term preservation for electronic literature. [1] The issues discussed in relation to the collection and preservation of electronic literature included (among others): selection, collection, context, intellectual property rights, and technical issues, the most crucial aspects of which will be presented in the following.

What is electronic literature?

“Electronic literature” is a simple-sounding label for a broad and multifaceted literary field that has evolved from as well as together with the internet. According to the U.S. Electronic Literature Organization, the term refers to “works with important literary aspects that take advantage of the capabilities and contexts provided by the stand-alone or networked computer” [2]. This includes two aspects: A technical and a sociological/communicational aspect. Apart from integrating the technological opportunities which the internet provides, electronic literature makes use of the particular interactive communication structure of the internet.

Electronic literature has (in the most cases) emancipated from the static, linear text narration to which print literature is bound. With the capabilities of the internet, innovative literary forms have developed in recent years: e.g. hypertext fiction, interactive and collaborative fiction, digital and audio-visual poetry, poetry that is generated by computers, or readable computer art installations.

The Electronic Literature Directory [3] introduced for browsing purposes a twofold genre classification: *Genre/Length vs. Technique/Genre*. While the former lists the relatively traditional categories Poetry, Fiction, Drama, Nonfiction, the latter distinguishes between the internet-specific literary forms Hypertext, Reader

Collaboration, Other Interaction, Recorded Reading/Performance, Animated Text, and Other Audio/Video/Animation, Prominent Graphics, Generated Text.

What to preserve? – Selection

The examples given in the paragraph above illustrate internet's primary forms of literature. There is no doubt that such works – like other contemporary literature, too – are worth preserving. Besides them, new forms of secondary literature have developed, for example literary journals, listservs, blogs, and wikis, which digital poets utilize in order to organize their community, but also to facilitate collaborative writing processes. Such secondary literature provides the philological context in which electronic literature is being created. It must be preserved to enable future literary studies. Both primary and secondary literature is in the focus of libraries as well as literary archives.

Another category of material is mainly of interest for literary archives and museums: writers' correspondences, manuscripts, personal diaries and calendars are collected and stored for future literary studies. Here, the digital world poses some particular challenges: Instead of or in parallel to a journal, the internet poet may write a blog on his homepage. His correspondence is stored, for example, in his Outlook folder on his hard disk, and so is his calendar. His manuscript may consist of a number of Word files, the newest version always overwriting the previous ones. Consequently, tracking down changes may be impossible.

So, while there is a plentifulness of chat and notes on webpages, and archives can hardly manage to capture all of it, the long-term preservation of relevant resources is seriously endangered.

Libraries, archives and museums need strategies how to cope with this dilemma. They are faced with the question of whether they have the capability to manage and the means to afford the preservation of all of those literature related resources that can be found online. The German Literature Archive and the German National Library are still refining their selection profiles with regard to electronic online literature. While each institution will need to define its own specific selection profile, it may also be feasible for the community as a whole to agree on certain common selection criteria.

Apart from selection criteria for electronic literature, literary archives need a policy how to deal with the new kind of "literary estates" that comes in on hard disks, CD-ROMs and USB-Sticks instead of boxes full of inscribed paper.

Harvesting the literary web – Collection

In order to guarantee long term availability of electronic literature resources, libraries and archives store local copies on their servers, respectively repositories. To this end, the selected resources can be harvested from the internet once or in periodic intervals. Archives and libraries can already choose among a range of existing software solutions. According to a predefined harvesting

policy, a list of predefined URLs is downloaded to the library's or archive's server automatically.

In contrast to printed works, where copies of published editions are collected, the collection of electronic literature requires new agreements. The definition of "edition" is challenged by the technological progress. The conventions of the print era can most likely be applied to electronic journals, which are usually issued periodically. Every new, completed issue can be collected.

It becomes more complicated with independent literary works, when borders between different editions are blurred because a version is constantly changed, refreshed, complemented etc. Where does one draw a line between different versions with regard to form and content, where does a new edition begin?

The same applies to authors' homepages, which are regularly updated, and to ever active blogs. Shall every change be reported and each revised version be transformed into an archival object? Shall objects be harvested weekly? Monthly? Daily? Shall old versions be overwritten or preserved together with the latest version? With regard to such questions, archivists and librarians, if necessary together with authors, need to reach sensible agreements.

At the Frankfurt workshop, the authors agreed with the archivists and librarians on a (rather traditional) approach to collect authors' homepages and primary works, literary magazines and dialogue forms like blogs.

The German Literature Archive already collects, indexes, and stores such material in a pilot operation. To this end, it cooperates with the state library service centre of Baden-Württemberg (BSZ) and shares the BSZ Online Archive. During the pilot phase, the German Literature Archive contacted the rights owners of selected literary resources and asked for their permission for harvesting their sites.

Only those sites for which permission was received were then downloaded to the BSZ Online Archive, indexed, and made accessible via the Online Catalogue of the German Literature Archive.

deutsches literatur archiv marbach	
Medienart	[Zeitschrift Internetquelle]
Titel	Electronic Journal Literatur primär / Hrsg.: Franz Kraiberger
Weitere Titel	e journal Literatur primär
Person	Kraiberger, Franz [Herausgeber]
Impressum	Wien
Ersch./Verlauf	1996 -
Fußnoten	"Das Electronic Journal bietet Essays und theoretische Beiträge zum Thema Neue Mediale Kommunikation an, ist jedoch auch offen gegenüber allgemeinen künstlerischen, literarischen Äußerungen, Absichten, Einsichten und Konzepten." [Information des Anbieters] Die seit 15 Zeitschrift enthält Beiträge vorwiegend österreichischer Literatur, Besprechungen und umfängliche Linklisten
Sprache	Deutsch
Land	Österreich
ISSN	1026-0293
Kette	S10_10_1_3_2 / Deutsche Literatur 1990 - / Einzelne literarische Zeitschriften
URL	http://www.ejournal.at/ http://ejournal.thing.at/ http://maxos.bsz-bw.de/boa/Dial/downloads/frei/78/0/index.html
Bemerkung zur URL	Verlag Verlag Langzeitarchivierung

Snapshot of a catalogue entry of an electronic journal at the German Literature Archive with reference to the original source on the internet and to the archival copy at BSZ Online Archive.

The collection procedure of the German National Library is not limited to electronic literature, but includes it. The library's collection field of online publications covers all text, image and sound-based works which are made available in public networks.

Suitable procedures for a large-scale collection, cataloguing and preservation of online publications are gradually developed and continually tested. For the time being, the German National Library uses an automatic collection procedure, which involves submission via registration form or OAI harvesting. While OAI harvesting is mostly used by large publishers, submission via registration is feasible for small publishers or independent authors, who release a manageable number of individual works online. It is probably the preferred method for electronic literature. When delivering via registration form, the author or publisher registers his online publication via a web form at the German National Library and manually delivers his publication to the library's deposit server.

Online publications are then catalogued by library staff, documented in the German National Bibliography and made accessible via the Online Catalogue of the German National Library. Persistent identifiers (URNs) ensure permanent addressing and long-term citability of online publications.

KATALOG DER DEUTSCHEN NATIONALBIBLIOTHEK

[← Zurück zur Trefferliste](#)

Treffer 10 von 154



Link zu dieser Seite	http://d-nb.info/975467271
Titel	Hypertextual fiction on the Internet: a structural and narratological analysis [Elektronische Ressource] / vorgelegt von Roman Zenner
Verfasser	Zenner, Roman
Erscheinungsjahr	2005
Umfang/Format	Online-Ressource
Anmerkungen	Langzeitarchivierung gewährleistet
Hochschulschrift	Aachen Techn. Hochsch., Diss., 2005
Persistent Identifier	urn:nbn:de:hbz:82-opus-11023
URL	Archivserver der Deutschen Nationalbibliothek: http://sylvester.bth.rwth-aachen.de/diss... http://sylvester.bth.rwth-aachen.de/diss...kostenfrei
Schlagwörter	Internetliteratur ; Erzähltheorie ; Online-Publikation
DDC-Notation	802.85 [DDC22ger]
Sachgruppe	800 Literatur, Rhetorik, Literaturwissenschaft

Snapshot of a catalogue entry of an electronic resource at the German National Library with reference to the original source on the internet and to the archival copy at the library's archival server.

Context

The creation of electronic literature involves hard- and software, a number of applications, and a set of technologies. The preservation of electronic literature implies the preservation of all of these technical components, so to speak of the technical "environment" under which it was created – or the replacement of this technical "environment", for example per emulation.

A similar challenge is to preserve the sociological environment of an electronic document; this means the context in which an electronic text is embedded. Electronic documents may for example refer to other documents via hyperlinks, to images embedded in the text, to comments added to it etc. When collecting electronic literature, a decision has to be made, how much of a text's context ought to be captured.

Experiences with harvesting tools reveal that it is not easy to capture the context, in which an electronic document was originally published. The most obvious problem: It is in the nature of selective harvesting that only the desired pages are harvested. Consequently, external links are deactivated. So the original "environment" of a web page is lost. Another problem consists of externally generated elements, like images, a calendar function or even advertisement.

To preserve dynamic elements is another specific challenge. Of all things, dynamic is a constitutive element of electronic literature. Beat Suter introduced a model of the "development of electronic writing" [4], in which he distinguishes four phases of electronic writing, marked by increasing use of dynamic and interactive elements:

1. flexible text
2. hypertext
3. networked writing
4. "pending writing"

Flexible text means linear text that is simply generated on a computer (e.g. an electronic manuscript that is used for generating the printer's copy). The flexibility is characterised by the opportunity to "cut and paste" text blocks, to move lengthy passages within one document or among different documents. The publishing format for flexible text is typically PDF.

Hypertext is as well first generated on a computer, but afterwards converted into HTML and made available via the internet. Hypertext utilizes hyperlinks, which interconnect a number of text fragments. Hypertext is no longer linear, because the reader can browse through the text fragments on individual pathways. Moreover, the author has the possibility to link from his text to external references, thus establishing intertextual references.

The next stage, networked writing, implies dynamic or interactive features, like commenting and collaborative writing processes. Typically, several authors team up for a networked writing project and share a mutual working space, often in public. So the text is made available to readers well ahead of its completion. The reader can participate in the writing process by commenting or by getting involved in discussions with the authors.

"Pending writing" is the most elaborate form of current electronic literature. It is characterized by a dynamic, interactive writing process and the use of various technological features such as computer-assisted text generation or mechanical, arbitrary organization of text. "Pending" works are always in a state of incompleteness, because authors and readers constantly interfere with the text, modify or manipulate it, add own texts or links to external resources. Thus, the context itself becomes a constituting feature of a pending text.

Technical Issues

The technical complexity of preserving electronic literature increases with each phase described above. Objects of the first two phases – text files and hypertext – can relatively easily be preserved as single objects. The third phase, networked writing, produces multilayered objects. For preservation, such complex objects have to be fragmented into a number of single objects. Emulation appears to be a workable solution, too. The preservation of objects from phase four, pending writing, requires very elaborate emulators.

The fundamental claim of the German Literature Archive and the German National Library is to document and preserve the digital avant-garde literature as comprehensively as possible. This implies that on the one hand neither of the two institutions wants to exclude any data formats from its collections. On the other hand, with an increasing amount of archived formats, the complexity of preservation measures increases manifoldly.

The participants of the Frankfurt workshop agreed that the best way seems to be to involve authors in the preservation process of their more complex works in order to document the entire compilation environment. Especially in order to successfully and adequately preserve the more complex networked works and dynamic forms of the “pending writing” process, the collaboration of authors, archivists, and librarians appears necessary.

Intellectual Property Rights

Authors and archivists as well as librarians are troubled by many unresolved legal questions relating to the preservation of electronic literature: Under what circumstances are archives and libraries allowed to create copies of the archived objects? Must copies be authorised by authors? Is the owner of a literary blog allowed to grant intellectual property rights on all entries in his blog to the archive? What about the rights of third parties like web designers?

The German National Library collects electronic literature under national legal deposit legislation. It does not have to request the right holders’ consent before harvesting their websites. The German Literature Archive and other archiving projects are faced with the necessity to ask permission every single time they want to collect a new web resource.

The fact that national legal deposit libraries are allowed to collect web resources does not mean that they are entitled to make them automatically available. Like other libraries, the German National Library has to negotiate access conditions individually with the right holders.

In order to simplify the resolution of such intellectual copyright issues, the participants of the Frankfurt workshop advocated the adoption of public licences for electronic literature, of which Creative Commons might be the most widely known.

Perspective

In a field as multifaceted as electronic literature, the involvement of authors in preservation processes seems strongly advisable. The potential synergy that lies in a common consent with regard to selection criteria, the process of collection building, legal solutions, and the technical framework, to mention the most crucial aspects, increases the likelihood that today’s electronic literature will be preserved for future generations. The Frankfurt workshop can be seen as a first step in the right direction of such collaboration for the German-speaking area. The participants are determined not only to continue but even to extend their collaboration: Further steps are envisaged, such as the compilation of a “preservation guide for authors”.

References

- [1] Documentation of the workshop is available from URL:
<http://www.langzeitarchivierung.de/modules.php?op=modload&name=Downloads&file=index&req=viewswndload&sid=29>
- [2] Electronic Literature Organization. URL:
<http://eliterature.org/about/>
- [3] Electronic Literature Directory. URL:
<http://directory.eliterature.org/index.php>
- [4] Beat Suter: Das Neue Schreiben 1.0. (2004) URL:
<http://www.netzliteratur.net/suter/dasneueschreiben1.html>

Preservation of Art in the Digital Realm

Tim Au Yeung, Sheelagh Carpendale and Saul Greenberg

University of Calgary
2500 University Drive North West
Calgary, Alberta, Canada
ytau@ucalgary.ca, sheelagh@ucalgary.ca, saul.greenberg@ucalgary.ca

Abstract

This paper discusses the challenges of preserving art in the digital context. It provides an overview of the broader digital preservation challenge, and then considers new media art within that context. Through several case studies, it illustrates and discusses problems, issues and proposed solutions to digital art preservation. We will see that while work has been done towards digital preservation, significant issues remain.

Preservation of Art in the Digital Realm

The preservation of information is the cornerstone of human progress – by passing knowledge from one generation to the next using a multitude of symbols, devices, tools and approaches, civilization has been able to advance. Throughout history, art has played an important role in this transmission with artistic depictions being more than representations of the world but reinterpretations for the sake of communicating what is deemed important. It is critical to note that these reinterpretations reflect not just the material culture but how society understood its place in the universe. Their understanding of the world comes to us largely from surviving artifacts including many art objects. Cultural heritage institutions like museums, archives and libraries have taken custodianship of these artifacts for the sake of the preservation of knowledge. In doing so, the exercise has become institutionalized with both the practices and the policies for collecting becoming formalized. At the same time, the institutionalization has led to a smaller number of individuals able to engage in a discourse on the values, implications and impact of choices made in knowledge preservation, to the point where the domain is primarily composed of specialists.

Recently, the transformation of society into a networked digital culture with millions of creator-publishers is eroding the underpinnings of institutionalized knowledge preservation and creating a challenging environment to preserve modern culture. This paper will explore the issues in the preservation of art in the digital realm both from the context of institutions and creators. It will begin by examining the broader digital preservation context before narrowing to the preservation of art.

The Broader Context of Digital Preservation

Introducing Digital Preservation

In one of the seminal works on digital preservation, *Preserving Digital Information* (Waters & Garrett, 1996), the authors observe that “the first electronic mail message was sent from either the Massachusetts Institute of Technology, the Carnegie Institute of Technology or Cambridge University. The message does not survive, however, and so there is no documentary record to determine which group sent the pathbreaking message.” Such events are all too frequent in the history of digital information and reflect its ephemeral nature. This also emphasizes that it is not simply a technical problem but “[r]ather, it is a grander problem of organizing ourselves over time and as a society to maneuver effectively in a digital landscape.”

Much as cultural heritage institutions hold physical artifacts, the report identifies the basic unit of preservation in the digital context (the information object) and notes at least five aspects that impact the integrity of the information object: content, fixity, reference, provenance and context. The type of *content* the information object is can determine the kinds of activity necessary to preserve the information object. The *fixity* of an information object identifies issues related to the dynamic nature of digital information and how to address incremental versions. How one *references* an information object impacts its integrity in terms of locating it. One particular challenge here is that information objects can be located in many places leading to the question of the authoritative version of the object. The issue of authority and authenticity of the object directly ties to the issue of *provenance*. Where the chain of custody for a physical object must be singular, it is not so for digital objects. This is important because when the object is changed, it becomes much harder to determine whether the change is an authentic change (as coming from a source with the authority to make the change) or a spurious change (coming either from malicious intent or inadvertent corruption). Finally the *context* of an information object has impact on its preservation and includes the technical context for viewing as well as related and supporting objects.

Why is Digital Preservation Harder?

All of the above apply to physical as well as digital objects. This raises the question for the difference between the physical and the digital and particularly why digital is more difficult. The Digital Preservation Coalition (Jones and Beagrie, 2002) note a number of factors for why digital preservation is harder. Machine dependency, speed of change, fragility of media, ease of making changes and the need to make changes, the need for active preservation and the nature of technology all play into making digital preservation harder than traditional preservation.

In terms of *machine dependency*, the fact that one requires an intermediary means that preservation work can only be assessed in the context of the original viewing environment. If that viewing environment is unavailable, then there can be no certainty that what one is currently viewing is reflective of the original. The *speed of changes* is also significantly greater than in traditional media. Where the shift from stone to paper reflects a shift over thousands of years, the shift from punch cards to optical media reflects a matter of decades. Similarly, digital media is physically *fragile* and requires a supporting technology infrastructure. *Ease of change* coupled with the need for *active preservation* raise the spectre of repeated inadvertent changes over the life of an object – changes that can corrupt and alter meaning. The need for active preservation is directly related to the speed of change and the fragility of the material where the tradition of benign neglect that for the most part worked effectively with traditional preservation will not work with digital.

The final reason why digital preservation is harder is simply that it is unknown at this point. While the conservation of analog materials is a well-known exercise, digital preservation practises are still largely untested. Recommendations are not necessarily supported by evidence and can be costly to implement.

Problems with Digital Preservation

Moving beyond the basic enunciation of the problem of preserving digital information objects to exploring the specific challenges associated with digital preservation, Besser (Besser, 2000) notes that there are five specific problems related to the preservation of digital information. The first problem is the *viewing problem*, which relates to the technical context noted above. The naked human eye can view physical artifacts but digital objects require technology to be viewed. A second problem is that of *scrambling* where the digital object may have an additional layer of complexity added to it through compression to save space or through encryption either due to security concerns or because of copyright management issues. The third problem in Besser's ontology is the problem of *inter-relation*. With traditional media, the object tends to be a singular, discrete item. With digital objects, it can be a conceptual construct composed of many individual digital files. The fourth problem is the *custodial problem* – this being directly related to provenance above. Where

institutions have divided the landscape of preserving analog material on a well-organized basis, no such divisions exist in the digital realm with organizations holding only part of what should be a coherent whole. Finally, there is the problem of *translation* where rapid obsolescence of file formats and digital standards results in digital objects being moved from one format to another to avoid obsolescence. However, the transformation from one format to another can cause the loss of information.

Approaches to Digital Preservation

While there are many subtle variations on a theme, there are typically six methods identified to address the problems of digital preservation listed above (NINCH, 2002). The first method is *technology preservation* and involves trying to save the actual environment required to view a digital object. This may involve saving the actual hardware and software and placing it in the environment where it can be maintained, often at a substantial cost. A second method is *technology emulation* where a substitute is developed for the original viewing technology. There have been questions about the practicality (Besser, 2000) but experiments have successfully demonstrated emulation (Seeing Double, 2004). *Data migration* is seen as an alternative to emulation, where the digital object is updated to run with modern software and hardware. However migration may not produce a perfect translation of the original and requires validation. Efforts like the Global Digital Format Registry attempt to make the validation process more efficient by providing a resource for centralizing knowledge on formats and best practises.

The first three methods are often seen as mutually exclusive but the next three are more supplementary, required regardless of overall strategy. First is *enduring care*, a catch all for activities necessary for good stewardship including recording keeping, safe storage and periodic checks. The second is *refreshing*, where new media periodically replaces the current medium to ensure the survival of the bits. Finally, the *digital archaeology* method involves reverse engineering to recover data from outdated and/or corrupted files and media.

Elements of a Digital Preservation Strategy

Regardless of the specific methodology used to preserve the digital objects, there are a number of elements of an overall digital preservation strategy that are consistently identified. Good metadata, trusted repositories, persistent identification, standards and best practices for handling, redundant storage and careful selection are all elements of a preservation strategy (Grout et al, 2000, RLG, 2002).

In the area of metadata and particularly *preservation metadata*, the institutional community has come out strongly for the need for metadata in preservation efforts. The belief is that metadata is necessary for the management and control of digital objects and the interpretation of the structure and content of the digital objects (Cedars, 2002). In specific, the PREMIS working

group refers to preservation metadata as “the information a repository uses to support the digital preservation process” (PREMIS, 2005). In fact PREMIS is the standard most cultural heritage institutions use as their preservation metadata standard. However, unclear best practices, a lack of support at the software level and uncertainty in the value of metadata impede adoption.

Similarly there has been a push towards *standardization of practices and formats* to simplify the problem. The belief is that if we use fewer formats and implementations, better tools and more unified techniques can be employed. This belief has spawned a number of best practice guidelines (NINCH, 2002, Grout et al, 2000, Jones and Beagrie, 2002) and projects like the Harvard Global Digital Format Registry and PRONOM. However creators tend to be unaware of these practices, requiring repositories to do the standardization (DeMulder, 2005). This approach adds to the cost of accepting materials and may result in repositories not accepting materials due to cost.

One goal of standardization and common practices is *distributed digital preservation*. In the simplest form this is redundant storage with most guides recommending two copies using different media. A more complex form of the idea is true distributed storage through a system like LOCKSS (Reich and Rosenthal, 2001) where organizations cooperate to store multiple copies.

Beyond secure storage there is the idea of defining the exact role of a repository to identify the characteristics of a digital institution that would reflect what a library, museum or archive represent with physical holdings. The formal definition of a *trusted digital repository* is “one whose mission is to provide reliable, long-term access to managed digital resources to its designated community, now and in the future” (RLG, 2002). This definition implies a number of things: that the institution goes beyond simply storing to managing the digital objects in its care and that the institution is situated within a community from which it draws its mandate and the specific means by which it preserves its objects. Most importantly, the goal is to provide access – this meaning that part of the mandate involves creating the tools by which viewers are able to interact with the digital object within future contexts.

Discussion of trusted digital repositories goes hand in hand with a discussion of the Open Archival Information System (OAIS, 2002). This reference model, originating from NASA, has been broadly adopted by the digital preservation community as a way of identifying the key characteristics of a preservation system. One of the most important aspects of the OAIS model is that it provides a common language and a common framework to discuss issues related to digital preservation.

The final element that is often emphasized is the issue of *selection*. Most best practice guides emphasize that the foundation for establishing a good digital collections rests on policies of selection and collection development. As one guide notes: “collection management policies that address digital materials, present the most critical challenge libraries or archives have ever had to face as

custodians of our scholarly and cultural heritage” (Cedars, 2002). While this is applicable to physical collections, the speed of change and loss has altered the nature of the role of the curator. From being passive receivers of cultural heritage, they have shifted to an increasingly active role where Eastwood observes that “[the] archival experience suggests that anyone responsible to select and preserve digital objects as records will have to seek materials actively in the here and now and be prepared to educate creators of them about the needs of long-term preservation” (Eastwood, 2004).

Digital Preservation in the Context of Art

Introducing the Art Problem

To open the discussion of the challenges facing the preservation of digital art, consider two largely positivist views of new media art conservation. Baker’s discussion on the symposium in January 2008 at the Getty Center titled “The Object in Transition” holds the role of conservators in a highly positive light. Baker outlines the extraordinary measures allocated to preserving the work “Indigo Blue” by Ann Hamilton (Baker, 2008), a work that crosses the line between sculpture, performance and process art. In both the work of the San Francisco Museum of Modern Art and in the discussions from the symposium, Baker reflects on the great efforts conservators expend on preserving works of art (like those of Eva Hesse) and their devotion to ensuring the survival of these pieces to future generations. A subtext one can take away from this discussion is that museum conservators would likely expend the same effort on the preservation of digital art.

Rinehart’s provocatively titled piece “The Straw that Broke the Museum’s Back?” echoes similar positivist views on the preservation of art, despite the title. His conclusion implies museums will succeed in preserving at least some digital art when he suggests that “[n]or are contemporary net.artists, working in undeniably ephemeral and center-less spaces, preventing the grand urge to collect, classify, and preserve” (Rinehart, 2000). In Rinehart’s vision of the future, museums and artists will collaborate in intimate fashion from the inception of the piece to its final form, documenting and making joint decisions on how the piece will continue to materialize in the future. Rinehart suggests the existence of solutions to the problem of preserving digital art is not risible but in fact entirely tractable through concerted effort and careful but early steps. It is worth noting that while these positivist views imbue the conservator with a great deal of credit (and resources), the reality is rarely so. As noted in Baker, the Berkeley museum did not have the resources to conserve Hesse’s “Auguht” and in fact few museums have the resources. A more pragmatic view comes from Besser (Besser, 2001). Returning to the digital preservation problem taxonomy, there are two problems specifically germane to electronic art – the problems of inter-relation and translation. Regarding inter-relation, web art is

challenging because the work often include references to web pages and sites central to the work but may not part of the work itself. If these pages change, the work itself may change in undesirable ways. In direct contrast to Rinehart's positivist take on the challenge, Besser fears that the "task may prove to be huge (and possibly intractable)".

Secondly, translation is problematic in that while digital art can be portable to different devices and contexts, these new contexts may alter the meaning. For example, consider Gary Hill's work where the work is meant for CRT displays and Hill's insistence that displaying the work on LCD flat-panel displays would be an alteration in violation to the spirit of the original work.

Besser goes on to identify characteristics of electronic art that make the problem different from the problem of analog or physical art works without electronic elements. In contrast to physical art, electronic art:

1. Lacks fixity
2. Can be dynamic
3. May have boundaries that are difficult to discern
4. May have critical format elements that make them challenging to work with but by changing them alter the work itself
5. May have difficulties guaranteeing authenticity
6. Can be malleable
7. Most importantly, can be difficult to define the precise nature of the work.

Besser poses the last characteristic in the form of the question "[w]hat really is the work?" and points to a 1980 piece "Hole in Space" that was simply a video feed between New York City and Los Angeles. If recreated, would this represent the work accurately? Would replaying the feeds from the time the installation stood from both NYC and LA be a sufficient representation of the work?

In placing digital preservation into the context of art, it is important to recap three trends evidenced by the broader digital preservation community.

1. The emphasis of digital preservation efforts has primarily been at the organizational level. In essence, digital preservation is an institutional effort that reflects institution priorities and resources. Selection and management policies are based on the challenges and goals of the institution.
2. The focus has been on the idea of the object – that it is possible to identify a discrete item. Discussions of information packages and bit-streams emphasize portability and manageability. The idea that it is possible to manage an object through its lifecycle also assumes discrete and concrete stages through which an object moves.
3. The goal has been towards standardization. The digital preservation community is heavily rooted in standards and best practices. Guides on best practices emphasize careful consideration to the kinds of material included in a repository and experimental work and prototypes often reflect the goal of moving incoming material into

"archival" formats that can be more easily handled, as they are better known.

These trends have a significant impact at the intersection of art and digital preservation and need to be explored to understand the particular challenges of preserving art in digital form.

The Notion of the Object in Digital Art

The question of the amorphousness of digital art raised by Besser is passionately argued by Jon Ippolito (Ippolito, 2004). He suggests that the fixity of the object endangers digital art itself, that "[w]hile the reductionism of the wall label enfeebles conceptual and single-performance art, it threatens to obliterate digital culture completely." Instead, he argues "new media artwork must keep moving to survive". Ippolito points to a number of dimensions where new media art breaks out the traditional bounds that conservators would like to place on the work. He suggests that unlike traditional art, new media art has variable authorship, titles, dates, media, dimensions and even collections.

For instance Winget (Winget, 2005) describes the piece "Loops", a portrait of Merce Cunningham by Paul Kaiser, Shelly Eshkar and Marc Downie. The piece combines sensors on Cunningham's hands to record the movement from Cunningham's "Solo Dance for Hands and Fingers", which is then interpreted by an artificial intelligence algorithm to display the sensor nodes in conjunction with recorded narration and music. However, not only do these work in conjunction with one another but the piece also changes in the presence of viewers. As with the piece "Hole in Space", it raises the question of what to preserve. As Winget notes, videotaping any given instance is incomplete and unlikely to capture the essence of the piece, but if you have to restage the piece, one is left with questions as to what are the essential features of the piece that need to be restaged and what features can be altered to reflect the changes in the technical environment.

The Institutionalization of Art

Issues surrounding of the institutional nature of art and in particular art conservation and preservation are not endemic to digital art. This is a challenge across all genres of art. In particular, the co-mingling of artists and conservators at earlier and earlier stages of the work raises questions as to the nature of that institutionalization. So when Rinehart (Rinehart, 2000) calls on the art community to define the types of metadata required and to develop methods for intellectual access to digital art, to which community is the question addressed to? Is there actually a cohesive organization that can speak for artists across all genres and types to answer these kinds of questions? Clearly this is a rhetorical question as there is indeed no singular entity that can address issues for all artists – there are both many organizations and there are no organizations where independent artists are concerned. Yet unless artists undertake the role of preservation themselves, the

decisions as to what to collect and how to preserve will rest in the hands of institutions and organizations potentially without regard to the sensitivities of the artists.

In particular, new media and digital art tends to be subversive in nature, bucking the general paradigm espoused by the prevailing institutions that reflect normative identity and majority views. As Lloyd notes (Lloyd, 2007), this is problematic as “[d]ecision makers do not have the resources to preserve everything. Therefore, decisions have to be made about what is significant, and, consequently, whose interests are to be acknowledged, what documented history is to be privileged, and whose history is to be marginalized or silenced.” While Lloyd is speaking towards cultural heritage materials, this idea of significance can certainly be extended to art. This is especially so in the case of digital art where intervention must occur early and often. In such cases, conservators and decision makers may not have the benefit of hindsight to identify works of cultural significance and the act of collection and preservation may pick winners and losers in the game as it were.

This, however, assumes the hegemony of the institution in the preservation of digital art. Gracy (Gracy, 2007) would argue that another possibility is a more likely reality: that “the curatorial or archival authority with which cultural heritage institutions are invested may diminish to the point where society may question the need for such entities to perform such work” as technologies of disintermediation become more widely available. With websites like Flickr and YouTube, individual viewers have the ability to curate their own collections and act in ways necessary to preserve the work. This comes as little surprise to new media artists as the community has been outside of the mainstream for some time and has experimented with alternative approaches to curating work. Grubinger’s experiment with C@C (Grubinger, 2006) was ground breaking in allowing artists to curate other artists’ work as part of the overall interaction process. While the experiment was ultimately abandoned, it can be argued that the idea was simply before its time. Later projects like low-fi and turbulence have taken up the banner of independent curation. Paul notes that “even though it may not be their explicit goal, these projects implicitly challenge the structures of legitimation created by the museum system and traditional art world” (Paul, 2006). Instead Paul sees the reconfiguration of the traditional roles of the curator, artist, audience and museum due to the transformative nature of the technology, technology that allows distributed curation, automated filtering by software and wider dissemination of works than at any other prior time.

Here then is the contradictory challenge of institutions in the context of digital preservation. On one hand, institutions may act in a pre-emptive manner selecting out some for wide dissemination and preservation while leaving others out not through the benefit of broader discourse on the value and meaning of the works but due to pragmatic matters reflective of individual institutions and policies, policies which may be out of date or incomplete.

On the other hand, the power of digital dissemination may reduce the legitimating role of institutions to the point where their value as entities comes into question. Yet, without institutions, preservation for the common good becomes problematic. If YouTube and Flickr are cited as the type of democratizing forces that allows greater numbers access to artists disenfranchised by the traditional art institution, then what are we to make of the fact that they are commercial entities whose sole goal is the enrichment of their shareholders and not beholden to any notion of public good or enduring value?

Standard Art?

The issue of standards in the context of art is an especially interesting discussion. As Grubinger notes, “[a]rtists often embrace new technologies as a means in itself rather than a means to an end; they tend to fool themselves by the seemingly limitless possibilities of new techniques” (Grubinger, 2006). Artists who have embraced new media and digital art are likely pushing the leading edge of technology where standards have yet to form and practises either do not exist or are untested. This is problematic as museums are unlikely to be equipped to address the new and potentially complex formats that the artists are using. As such, museums may be reluctant to work with the piece compared to a work whose components are better known, leading to artists pushing the envelope being marginalized. What may be somewhat more troubling for artists though is the idea that their work should be constructed with preservation in mind. In the preservation study of *Ars Electronica* (Becker et al, 2006), some of the work was intended to be ephemeral in nature and therefore the choice of technologies and formats reflected an insistence on the transient. If museums and art galleries begin to insist that works be done to standards of preservation in order to be accepted by the institution, it may preclude artists who either are unable to work with the standards for technical reasons or who have made a conscious decision to make the work ephemeral in nature.

Case Studies in Preservation of Art in the Digital Realm

While the theories and strategies for digital preservation and art are still evolving, it is important to note that the community has not stood still. There have been a number of projects related to the preservation of new media and digital. Below are highlighted two projects, each representing a prototype for a specific approach to digital preservation and art.

Seeing Double

One of the most interesting exercises in digital preservation experimentation was an exhibition hosted by the Guggenheim Museum in spring of 2004 titled “Seeing Double” (Seeing Double, 2004). The goal of the exhibition was to bring together the original new media works and try to use emulation (see Rothenberg, 1998 for a fuller discussion of emulation approaches) to reproduce and re-

interpret the work. It was hoped that presenting the two together would allow both experts and the layperson to “decide for themselves whether the re-creations capture the spirit of the originals”. The range of techniques used varied from the simple storage and redisplay in Cory Arcangel’s “I Shot Andy Warhol” piece (where the hacked hardware limited the options to the restaging and filming of Robert Morris’s “Site”), to the creation of a software emulator to recreate the environment for the code in Grahame Weinren and Roberta Friedman’s “The Erl King”.

The interviews with the artists reflecting on the emulation effort of the exhibition were particularly interesting, where the range of opinions spanned the spectrum of responses. Weinren and Friedman viewed the new emulation hardware and environment as merely the carrier. In essence the “apparatus is no more than what makes the interactivity possible, so a digital version of the piece, whatever equipment it runs on, will be exactly the same piece.” This differed from John F. Simon Jr. (“Color Panel”) who felt variations are simply part of the process. Morris, in reflecting on the filming of the restaging of his piece, felt the recreation was more about the director than it was about himself as an artist. Finally, Arcangel felt that the piece would lose meaning without the corresponding hardware. If it were redone in fifty years, he’d want the original hardware, but failing that, not to have the museum try to recreate the hardware but rather to give away the software so that individual viewers could play with the code in their own context.

Ars Electronica

While the Seeing Double project is more about experimenting, the Ars Electronica project focuses on information gathering. Ars Electronica is one of the world’s largest collections of digital art in the world (Becker et al, 2007) and comprises over 30,000 works with 3,000 new works per year. A joint effort between the Vienna University of Technology and the Ludwig Boltzmann Institute Media.Art.Research undertook a pilot project to preserve a portion of the collection by trying to capture both the intentions of the artist as well as the experience of the viewer. The PLANETS digital preservation planning process was used to assess the essential characteristics of the works to determine the best course of action within the preservation context. By using workshops with interested parties like curators, art historians, computer scientists, specialists and management, the characteristics of the works are identified. The next phase of the pilot project is to use the information to implement a preservation strategy and evaluate the results.

Strategies and Solutions for Art Preservation

The majority of the strategies for preserving digital art fall within the rubric of solutions proposed by the broader digital preservation community. However, there also exists work specifically focusing on digital art preservation.

As Depocas suggests (Depocas, 2002), without documentation we would be unaware of the majority of the

panoramas from the 19th century and in particular, their influence on the public. He then draws the parallel with new media art. For digital art, greater viewership and access increase the likelihood of the work being preserved for the future. As a result, documentation is critical to the survival of digital art as it increases the opportunities for access. One challenge is to update the principles of documentation to reflect new media works where measures like dimensions no longer apply. As Depocas suggests, digital art in particular lies at the intersection of physical art objects and art events where they have an instantiation that changes over time. One important argument for good documentation rests in the need to reinterpret the work from its original context to the current context so that the viewer is aware of how the work was intended to be.

An extension of the idea of documentation is the Media Art Notation System proposed by Richard Rinehart (Rinehart, 2007). Rinehart suggests that digital and media art forms have greater similarity to music than traditional visual art forms and suggests that how music is preserved and passed on can provide inspiration for how to document new media and digital art. What Rinehart proposes is a systematic approach for documenting media art so that it can be “played” back in different technical contexts but with end results as intended from the artist – in essence, a score for any performance of a new media piece. However, unlike musical scores which have a specific language that one must learn, Rinehart proposes couching the MANS system in an existing notation system, in this case XML, to reduce complexity and increase adoption.

Complex Media Art: An Example

While many new media projects involve some degree of technology, the issues of experiential pieces, emergent technologies and complex interaction are often most fully realized in projects developed between computer scientists and artists. One such case is a course co-jointly taught by the University of Calgary and the Alberta College of Art and Design. In this course, students drawn from computer science and art are given the task of jointly developing a piece that explores issues in both disciplines. The resultant pieces produced typically include software to control the piece, physical interaction and reactivity to the viewer.

In the most current iteration of the course (2008), pieces included: a video booth where the reactions of the viewer to pre-selected videos were recorded as a means of influencing the next viewer; a meditative piece involving projecting Persian patterns into a reflecting pool where the drawing of the pattern is influenced by the viewers around the pool accompanied by audio recordings of spoken Persian poetry; a large screen projection with 3D animations where the animations are determined by the presence and location of sculpted figures on a chess-like board; another 3D projection where the viewer can alter the perspective of the projection through a large button; and an interactive piece where viewers can draw using large virtual crayons onto a projected surface with the movement of the crayons generating tones.

Consider, for instance, the video reaction booth. The work consists of a telephone booth sized wooden box. On the side is a single computer monitor. Inside is a second monitor with a set of three buttons, a chair with a sensor mounted to it and a web camera. The monitor on the outside continuously loops still images of the recorded reactions of previous viewers. When a viewer enters the booth and sits down at the chair, their presence is signaled to the system where it starts recording (through the web camera) just the head of the viewer from a frontal perspective. This recording now also starts displaying as still images on the external monitor alongside the previously recorded streams. The viewer inside the booth is presented with an interface consisting of a gallery of pre-recorded video streams of the reactions of viewers to videos that range from extreme topics including car crashes and self-immolation videos to videos of laughing babies.

To analyze this work from Besser's typology, we have a number of issues. From the standpoint of the viewing problem, two research technology frameworks were used to create the display: *Phidgets* to provide physical user interfaces and *Processing* to handle video / on screen user interfaces. As each framework is based heavily in research activities, they lack the stability of commercial products. More importantly they have the potential for changing over time or being abandoned when the research value is no longer there. Since both frameworks are independent of the artwork, substantive changes to either framework could impact heavily the ability to restage or migrate the software driving the work. From the standpoint of interrelation issues, consider the dynamic nature of the work. As viewer reactions are recorded, the experience of the work changes for any subsequent viewer. A viewer encountering the work devoid of any recorded reactions will have a very different sense of the work compared to a viewer coming into the work with a large quantity of reactions recorded. Equally important, we have two viewer contexts to how the work is experienced – the outside experience and the inside experience. Scrambling is also an issue as video formats are invariably compressed to save space and improve performance. From a custodial perspective, the work represents a challenge in terms of the rights of those being recorded. Would transferring the work from one institution to another allow for the transferring of the recorded reactions? If not, those experiencing the work in the new location would be engaged in a new experience. Finally, the issue of translation would be problematic as there are two technical frameworks in addition to the base computer system and specialized hardware that would have to be translated from one instantiation of the work to another. Documentation would be critical to restaging the work but this is a case where even the documentation is complex. Because the work is the product of two people with very different aims (the artist and the computer scientist), assessing the aspects of the work that would be critical for restaging the work would depend entirely on whom you asked. All of this raises the question on whether the work could be preserved

in a way that future viewers could experience as intended or whether the documentation would exist solely to record the experience as it occurred.

Conclusion

While this paper does not provide any definitive answers as to how art and digital preservation will play out in the next twenty years, this is because that future is still quite murky. Programs like the NDIIPP in the US and PLANETS in the EU are attempting to address issues at a very broad level. Museums are still trying to shake the idiosyncratic nature of their heritage and collaborate in a networked fashion in ways that their library and archives brethren have long since adopted. Artists are just starting to explore the limits of digital technology. These are not questions that will be answered in the near future. However, what has been presented is a broad overview of possible directions. While work has been done to classify and identify the digital preservation issues, things like Besser's taxonomy are not substantively supported by empirical findings but reflect anecdotal observations. Solutions like migration and emulation still have to be tested against a large corpus of material beyond that of the current test sets. Even the durability of the physical carrier media is still in doubt with only good longevity tests having been done on magnetic tapes for data. The impact of the network and democratizing means of publishing have not been fully realized in the context of art nor have new economic models fully taken hold yet in the art world. This leaves in question where the resources for the preservation of digital and new media art will come from given that current institutions are stretched thin with existing challenges. Thus the lack of a definitive conclusion is a reflection of a field at a very early stage where much remains in flux.

References

- Baker, K. (2008, March 2, 2008). Saving the Soul of Art. *San Francisco Chronicle*.
- Becker, C., Kolar, G., Kung, J., & Rauber, A. (2007). *Preserving Interactive Multimedia Art: A Case Study in Preservation Planning*. Paper presented at the 10th International Conference of Asian Digital Libraries.
- Besser, H. (2000). Digital Longevity. In M. K. Sitts (Ed.), *Handbook for Digital Projects: A Management Tool for Preservation and Access*. Andover, Massachusetts: Northeast Document Conservation Center.
- Besser, H. (2001). *Longevity of Electronic Art*. Paper presented at the ICHIM 01: Cultural Heritage and Technologies in the Third Millennium.
- Cedars Project. (2002). *Cedars Guide to Digital Collection Management*. Cedars Project. Retrieved

- March 22, 2008 from <http://www.leeds.ac.uk/cedars/guideto/collmanagement/>.
- Cedars Project. (2002). *Cedars Guide to Preservation Metadata*. Cedars Project. Retrieved March 22, 2008 from <http://www.leeds.ac.uk/cedars/guideto/metadata/>.
- Depocas, Alain. (2002). *Digital Preservation: recording the recoding – the documentary strategy*. The Daniel Langlois Foundation. Retrieved March 24, 2008 from <http://www.fondation-langlois.org/flash/e/index.php?NumPage=152>.
- Depocas, A., Ippolito, J., & Jones, C. (2003). *Permanence Through Change: The Variable Media Approach*. New York: Guggenheim Museum Publications and The Daniel Langlois Foundation for Art, Science and Technology.
- DeMulder, Tom. (2005). *DSpace@Cambridge: Implementing Long-Term Digital Preservation*. Retrieved March 22, 2008 from http://www.dspace.cam.ac.uk/bitstream/1810/104791/1/Rosetta_Stone_paper.pdf.
- Eastwood, Terry. (2004). Appraising Digital Records for Long-Term Preservation. *Data Science Journal*, vol. 3, 30.
- Gracy, Karen F. (2007). Moving Image Preservation and Cultural Capital. *Library Trends*, 56.1. pp. 183-197.
- Grout, Catherine, Phill Purdy and Janine Rymer. (2000). *Creating Digital Resources for the Visual Arts: Standards and Good Practice*. Online: Visual Arts Data Service. Retrieved March 22, 2008 from http://vads.ahds.ac.uk/guides/creating_guide/contents.html.
- Grubinger, E. (2006). 'C@C': Computer-Aided Curating (1993-1995) Revisited. In J. Krysa (Ed.), *Curating Immateriality: The Work of the Curator in the Age of Network Systems*. New York: Autonomedia.
- Harris, J. (2007). Control, Alt, Delete? [Electronic Version]. *Mute*. Retrieved March 8, 2008 from <http://www.metamute.org/en/Control-Alt-Delete>.
- Ippolito, Jon. (2004). *Death by Wall Label*. Retrieved March 23, 2008 from [http://three.org/ippolito/writing/death_by_wall_label@m.html](http://three.org/ippolito/writing/death_by_wall_label@.html).
- Jones, Maggie and Neil Beagrie. (2002). *Preservation Management of Digital Materials: A Handbook*. Online: Digital Preservation Coalition. Retrieved March 22, 2008 from <http://www.dpconline.org/graphics/handbook/>.
- NINCH Working Group on Best Practices. (2002). *The NINCH Guide to Good Practice in the Digital Representation and Management of Cultural Heritage Materials*. Online: The National Initiative for a Networked Cultural Heritage. Retrieved March 22, 2008 from <http://www.nyu.edu/its/humanities/ninchguide/>.
- Ippolito, J. (2007). *Death by Wall Label*. Retrieved March 22, 2008, 2008, from http://three.org/ippolito/writing/death_by_wall_label@m.html
- Lloyd, Annemaree. (2007). Guarding Against Collective Amnesia? Making Significance Problematic: An Exploration of Issues. *Library Trends* 56.1, pp. 53-65.
- Paul, C. (2006). Flexible Contexts, Democratic Filtering and Computer-Aided Curating. In J. Krysa (Ed.), *Curating Immateriality: The Work of the Curator in the Age of Network Systems*. New York: Autonomedia.
- PREMIS Working Group. (2005). *Data Dictionary for Preservation Metadata*. OCLC. Retrieved March 22, 2008 from <http://www.oclc.org/research/projects/pmwg/premis-final.pdf>.
- Reference Model for an Open Archival Information System (OAIS)*. (2002). Retrieved March 22, 2008 from <http://public.ccsds.org/publications/archive/650x0b1.pdf>
- Reich, V. and Rosenthal, D.S.H. (2001). LOCKSS: A Permanent Web Publishing and Access System. *D-Lib Magazine*, vol. 7, issue 6. Retrieved March 22, 2008 from <http://webdoc.sub.gwdg.de/edoc/aw/dlib/dlib/june01/reich/06reich.html>.
- Rinehart, R. (2000). The Straw that Broke the Museum's Back? Collecting and Preserving Digital Media Art Works for the Next Century. *Switch*, 6(1).
- Rinehart, R. (2007). The Media Art Notation System. *Leonardo - Journal of the International Society for the Arts, Sciences and Technology*, 40(2), 181-187.
- RLG. (2002). *Trusted Digital Repositories: Attributes and Responsibilities*. Mountain View: Research Libraries Group.
- Rothenberg, J. (1998). *Avoiding Technological Quicksand: Finding A Viable Technical Foundation for Digital Preservation*. Online: Council on Library and Information Resources.
- Seeing Double: Emulation in Theory and Practice*. (2004). Retrieved March 22, 2008, 2008, from <http://www.variablemedia.net/e/seeingdouble/>
- Waters, D., & Garrett, J. (1996). *Preserving Digital Information: Final Report and Recommendations*: Commission on Preservation and Access and The Research Libraries Group.
- Winget, M. (2005). *Digital Preservation of New Media Art Through Exploration of Established Symbolic Representation Systems*. Paper presented at the JCDL 2005 Doctoral Consortium

In Cypher Writ, or New Made Idioms: Sustaining Digital Scholarship as Cooperative Digital Preservation

Bradley J. Daigle

University of Virginia
Alderman Library / PO Box 40155
Charlottesville, VA 22904 USA
bradley@virginia.edu

Abstract

Digital Scholarship is a method of scholarly communication, research, and exchange of ideas that employs modern forms of technology, in particular, those forms of technology maintained within an institution's cyberinfrastructure. Digital scholarship then is often, in equal parts, the intellectual content *and* the manner in which it is created and presented. That is what sets it apart from, for example, humanities scholarship as it has been historically undertaken in its published form. Thus it would follow that the sustaining of digital scholarship goes far beyond what is commonly known as digital preservation. In other words, sustaining digital scholarship is not just the difficult task of preserving the atomized digital objects (or even bits and bytes) but also the relationships among them. These relationships represent the digital world of authorial aggregation and distribution that also needs to be preserved. This is not a task that any one unit within a university can possibly undertake. This article provides an outline of activities that are taking place at the University of Virginia and provides some outlines and strategies for approaching such a complex problem set.

What is Digital Scholarship?

This book, as long-lived as the elements
Or as the world's form, this all-gravèd tome
In cypher writ, or new made idiom;
We for Love's clergy are only instruments;
When this book is made thus,
Should again the ravenous
Vandals and the Goths invade us,
Learning were safe; in this our universe,
Schools might learn sciences, spheres music, angels verse.
John Donne "Valediction to his Book"

Centuries after Donne, we are less confident than ever before that "Learning were safe." Libraries continue to struggle to preserve the bulk of materials that are familiar to most: books and paper. Some would argue that this front, at least, has been contained. What does digital

preservation mean with respect to today's digital technology? How are scholars taking advantage of new methodologies for doing what has always been the major product of higher education—research? With new trends and even newer avenues of technology to explore, the pressure mounts on academic infrastructure to continue to preserve the scholarly output of its faculty and students. Recent trends point to an understanding that a broader audience is needed to tease out the full implications of digital preservation. The Digital Preservation Coalition's report, *Mind the Gap: Assessing Digital Preservation Needs in the UK* undertaken in 2006, reiterates that it is critical that we broadcast this message to as wide an audience as possible.¹ Any complex set of preservation activities is rendered far more difficult in the wake of the digital revolution and for academics in particular, digital scholarship. It is clear that no one unit, or even no single institution can achieve this in a feat of individual prowess—the resources needed are too great and the scope too vast. Cooperative practices, ingrained and entrenched, are our only hope to succeed to preserve digital scholarship.

Digital Scholarship is the "new made idiom" for how many scholars now undertake and present their research. It is a

¹ The report highlights the following key elements which are worth reiterating here:

- Organisations should continue to raise awareness of the impact of digital preservation beyond the current core of informed individuals and institutions.
- Training in digital preservation should be encouraged and programmes should be integrated into the training of professionals such as conservators, librarians and archivists.
- Awareness of digital preservation issues should be raised at government level, both nationally and internationally, in order to influence relevant policy making.

An international collaborative 'market' for digital preservation tools should be created. Such a market should encourage the use of open file formats and standards and consider the long-term preservation needs of digital information.

<http://www.dpconline.org/graphics/reports/mindthegap.html>

relatively recent trend that has many libraries—in particular academic libraries—scrambling to develop the requisite service models to both support and sustain it. Digital Scholarship incorporates more and more digital media for research and classroom-based projects. It goes beyond the relatively straightforward landscape of electronic journals that originally were considered to be prototypical digital scholarship examples. I see digital scholarship as a method of scholarly communication, research, and exchange of ideas that employs modern forms of technology, in particular, those forms of technology maintained within an institution's cyberinfrastructure. The American Council of Learned Society's report on cyberinfrastructure entitled, *Our Cultural Heritage*, boldly indicates that the authors believe this form of scholarship is the future of *all* scholarship (ACLS 2006). In this essay I will be specifically addressing how digital scholarship taxes our notions of appropriate curation and digital preservation. In particular, I will be looking at practical approaches to developing services, infrastructure, and policy related to these activities.

How is Digital Scholarship Different?

In what manner is digital scholarship different from “traditional” scholarship? Donne's poem referenced above celebrates the book as a stable vehicle for the dissemination of “learning” in an age that witnessed the harbinger that was to become print culture as we know it. The transition from an oral to written culture and then from manuscript circulation to print production marked a shift in technology. The hegemony of the codex format is still very much with us for many good reasons that I need not detail. However, today, new forms of scholarship are available through the ubiquitous use of technology. Data can be mined, texts can be structured, images can be delivered and manipulated—all with some very basic tools. This is where the simple comparisons end. Digital Scholarship embarks into highly esoteric realms—realms that few may even know existed. New advances in computational science, data set manipulation, aggregation of digital objects all take on increased magnitudes of complexity.

These new planes of existence require ever changing and flexible architectures to manage and deliver this content. This takes us far beyond the realm of Donne's book and closer the digitally metaphysical. Digital scholarship then is often, in equal parts, the intellectual content *and* the manner in which it is created and presented. That is what sets it apart from, for example, humanities scholarship as it has been historically undertaken in its published form. Thus it would follow that the sustaining of digital scholarship goes far beyond what I would normally classify as the (already not so straightforward) preservation

of digital objects. In other words, sustaining digital scholarship is not just the difficult task of preserving the atomized digital objects (or even bits and bytes) but also the relationships among them. These relationships represent the digital world of authorial aggregation and distribution that also needs to be preserved. This is not a task that any one unit within a university can possibly undertake.

Core institutional services need to be developed in order to support and sustain digital scholarship in a manner that is appropriate to the institution's mission. These can be collecting strategies, organizational models, outreach services, as well as developing new tools for managing this scholarship. That said, digital scholarship requires a new form of library environment—one that is adaptable and extensible, one that properly adjusts to changing technologies. For most institutions this requires strategic partnerships both within and beyond what are often defined as traditional institutional relationships. I will later discuss what types of collaborative policies need be crafted. This will range from signed license agreements (SLAs) to collection or deposit agreements that cover the range of intellectual property and copyright issues. These policies should also detail how the work will actually be undertaken as it is a cooperative agreement between the author(s) and, in this case, the library as the future steward of the collection.

Goths and Vandals Invade?

When this book is made thus,
Should again the ravenous
Vandals and the Goths invade us,
Learning were safe

Like so many academic institutions, UVa Library struggles with the workload of managing and migrating legacy content along with the ubiquitous creation of new content. Digitizing activities are integrated in almost every facet of the higher education institutional framework both physically and philosophically. These voluminous activities threaten to strain the already tenuous hold libraries maintain over their digital services and support. One of the most important questions concerning the preservation of digital scholarship is: “How do scholars and librarians work together to ensure that resources created today will be available in the future?” (Marcum 2002). As we look at strategies for sustaining digital scholarship we are developing a framework for how all materials—old and new—can be properly stewarded. This has been a process I have been involved in here at UVa for several years. I hope to illustrate how we have begun to articulate the life cycle of digital objects (including their aggregate relationships) and how the sustaining of digital

scholarship is for us, the next generation of digital preservation.

What is a definition of Digital Preservation in this context?

Digital preservation is the managed activities for the long-term maintenance of a digital object and the continued accessibility of these objects. The Research Library Group defines digital preservation “as the managed activities necessary: 1) For the long-term maintenance of a byte stream (including metadata) sufficient to reproduce a suitable facsimile of the original document and 2) For the continued accessibility of the document contents through time and changing technology” (RLG 2002). It is also a practice that can simulate the original experience of digital scholarship (as I have defined it) whether that experience be approximated or emulated. No one would see digital preservation as a set of isolated activities in this context. It needs to form the core of any suite of services that are established in support of faculty and student research. We have demonstrated that we can easily create digital materials; we have yet to demonstrate that we can fully manage them. Digital preservation activities should move us “toward the realization that perpetuating digital materials over the long-term involves the observance of careful digital asset management practices diffused throughout the information life cycle. This in turn requires us to look at digital preservation not just as a mechanism for ensuring bit sequences created today can be rendered tomorrow, but as a process operating in concert with the full range of services supporting digital information environments, as well as the overarching economic, legal, and social contexts” (Lavoie and Dempsey 2004). Digital preservation requires and understanding of who own or is responsible for the scholarship.

At UVa we have adopted a managerial distinction to assist us in differentiating among all the possible players and preservation options. We have virtually (as opposed to physically) partitioned our service landscape according to who owns and/or manages the content that has been / will be created. We started with two main areas of content that allows us to build a structure that is both flexible and extensible. This content is defined as *scholar managed* or *library managed*. There are certainly more options but for our initial planning and development of a dependable cyberinfrastructure we began with these two categories. The scholar managed content environment is the product of supporting digital scholarship. It should be able to provide a faculty member with a stable suite of tools and services that will meet almost any need that arises through the creation of digital scholarship. Library managed content forms the core of the library’s digital collections and repository environment and includes content from our websites, databases, and OPAC. The library managed environment is our digital preservation infrastructure.

The purpose for virtually partitioning these two management spheres is that we wanted to create an integrated environment that allows users to crosswalk their content from scholar to library managed content arenas. In other words, the two partitions are based on very similar software platforms and identical hardware platforms. This way scholarship that a faculty member develops in the scholar-managed content environment already shares many of the basic hardware and software requirements for transitioning into the library managed area. If the scholar wants the library to preserve her digital scholarship we have a strategic *a priori* starting point.² By integrating these environments “behind the scenes” we hope to have much of the raw material that faculty need (and created by the library for faculty) managed in our environment and the researcher can draw up it in from the faculty environment. That way at least, the raw content has a preservation strategy (based on file types etc.) and the faculty member’s development is more related to the application and software functionality. This is our model for current and future scholarship. However, given the huge amount of legacy data the library manages, we have had to formulate a strategy for cross-walking much of the older material into the library managed content environment.

The Lay of the Service Landscape

In order to articulate the myriad of activities that comprise a strategy for digital preservation of this magnitude, we have broken out the entire process into several stages. What follows is a general overview of how we at the UVa Library approach this problem set. It is specific to our institutional landscape but by no means completely bounded by it. The goal in outlining the work plan is to allow others to adopt pieces or the entire process as a potential model for their home institutions.

We have two different vectors of approach for preserving digital scholarship. I classify them as *supporting* digital scholarship and *sustaining* digital scholarship. The former bespeaks of a highly collaborative, participatory role that librarians / technologists should have with faculty; the latter a set of transformative and migration activities with materials that have already been created and formed. Both

² This environment for faculty is meant to provide the “carrot” for using the system that the library has established in cooperation with several other university units. Faculty members can self-deposit in this environment but we make it clear that the faculty member manages the scholarship at this initial stage. For a good discussion of faculty self-deposit in IRs see Marshall’s article on the scholarly perspective, Section 4.

require a great deal of resources and planning and both are critical to the success of any institution's digital preservation strategy.

Supporting Digital Scholarship: Enhancing our Ability to Digitally Preserve

Activities that fall under this rubric can be categorized in many ways but most fall under support-service activities which can take the form of digital labs, digitization services, grant writing, and intellectual copyright consultation to name a few. Every institution has varying levels of infrastructure in place to support the teaching and research of faculty and students. These examples certainly represent an excellent beginning to a full suite of services for supporting scholarship. They enable the creation of new materials, their description, organization, and dissemination at a minimum. The services that are based on such activities draw heavily upon the expertise and abilities of both librarians and technologists (often in the form of blended professionals). In many cases, however, these services exist almost entirely independent of the second layer of support that is required—a complex institutional repository and web services environment. The maintenance of this framework often goes beyond a single department or unit's ability to support on its own. More and more, institutions are adopting the strategy of the institutional repository to administer faculty and student output. Foster and Gibbons see these types of systems as a form of digital preservation: "In the long run, we envision a system that, first and foremost, supports our faculty members' efforts to 'do their own work'" (Foster 2005). A recent survey of repository services demonstrates that very few (none with a Preserv³ profile) had a formal preservation policy (Hitchcock *et al.* 2007). Certainly, this is an important first step and the need to integrate the above-mentioned services with these repository environments is critical for truly supporting digital scholarship. It is a major part of the necessary cyberinfrastructure for faculty and student research. However, without the complete integration of services and repository environments it could still fall far short of a solid digital preservation solution. Too much of today's digital scholarship is taking place and exists only on faculty members' local machines which are managed informally and not part of an institution's infrastructure. This puts much of that work in peril for both the researcher and the institution: lose the scholar-lose the scholarship is not a sound institutional strategy. The organization of the repository landscape should represent the commitment of the library to preserve scholarly research as well as a concomitant assurance from the institution through its

support. If not, the result can be a series of one-off pseudo-solutions. Single solutions often address the preservation of files in isolation and are much less adaptable to aggregations of content. Integration of services and repository environments becomes part of what Lavoie calls an institution's promise to its scholars: "Fulfilling this promise requires the cultivation of stakeholder communities that, through their working and learning experiences, meaningfully engage with digital information environments" (Lavoie 2008). Cultivating these communities can occur in many ways—some overt and some that are covert. For example, most practitioners understand that in order to approach a solid preservation strategy one needs to "catch" faculty and students early on in the planning stages of their projects. At the very least, catching them at the point of production will minimize the efforts that may have to happen downstream whether they be reformatting, re-digitization, etc. These follow up activities can often derail future preservation strategies and damage relationships between the researcher and the institution.

Covert methods are often equally successful to those of a services lab or production environment. Creating an integrated environment that contains scholarship and projects is a key component. Ensuring that the faculty and students have a development environment that is built on similar standards (if not duplicative) that can be found in the institution's production and management environment will allow for smoother transitions between what I referred to as library managed content versus scholar managed content. There is of course the inevitable trade-off between standardization, which is essential for long-term preservation, and flexibility, which allows for a researcher's versatility in discovery and application. At UVa we have been collaborating for years with our central technology group, ITC (Information, Technology and Communications) to provide an appropriate technology environment that supports research. To create an environment such as the one needed to handle faculty scholarship the library could not do it alone.⁴ Instead, we built upon a relationship that centers on different spheres of management. In this scenario, the library is responsible for the content, ITC for the hardware. The software layer becomes the shared interface where a baseline platform is vetted and agreed upon. Producing a development environment that approximates the production environment is one way of approaching this problem. How would one

³ Preserv project <<http://preserv.eprints.org/>>.

⁴ This is a clustered server environment that provides three tiers of service: a development environment that individual faculty members can use to incubate their research and test out new technologies; a test environment that is a clone of the final production environment where changes and load testing occurs; and finally, the production environment which is meant to deliver and manage only fully tested and "mature" digital scholarship.

decide what new technology might need to be integrated into the production environment? Creating a feedback loop of testing and production can allow for greater flexibility. If a faculty member considers a new piece of software integral to her research then the library and other support structures have a review process to analyze and test the claim. If it is determined that the new technology provides new and improved functionality then the library can integrate the new technology into its environment along with the research. This provides the greatest balance between flexibility and stability. This is an ongoing cooperative approach to maintain a service environment that faculty use and trust. Beyond simply defining the environment (as if it were simple) the expectations that are required for the environment need to be clearly delineated. It took several months to establish service level agreements between the library and ITC in order that we could communicate those levels of service to faculty. For example, materials that are served from our production environment could have a 24 by 7 guaranteed “up-time” with a definable problem response time, the test environment might be 24 by 5, and then the development environment weekdays, 9 to 5. Establishing and publishing these parameters with faculty greatly increases the trust in the integrated environment and serves as an incentive for faculty to use our services to do their research rather than going it alone. Like Entlich and Buckley, we see it as our mission to create and “establish institutional repositories in which faculty are encouraged to deposit their work” (Entlich and Buckley 2006). If we do this, then preserving the materials becomes a slightly less difficult task since the cyberinfrastructure closely mirrors the library managed content environment.

Sustaining Digital Scholarship as the Next Level of Digital Preservation

Supporting and sustaining are not mutually exclusive activities. For larger institutions that were early adopters of digital technology, the support structures have changed dramatically over time. UVa is once such institution. Early activities originating from the mid 1990s to today mean that we have a vast amount of legacy data—none of which conforms to any one standard. Images, text, data sets, early faculty forays into digital scholarship, all sit on servers and laptops and any number of portable media devices. Enter the sustaining portion of digital scholarship. This is where I believe we push the limits of digital preservation. It often involves materials that used technology that has become obsolete or outdated file formats. The library is confronted with a series of challenges with this material. No single unit can make the decision to keep or weed the materials. Nadal speaks of a need of the need for the “human element” in digital preservation (Nadal 2007) and this

certainly comes into play in making these decisions. This is where the library needs to draw upon its collection development strategy for digital materials.⁵ At the very least this should provide some guidelines for prioritizing materials to be preserved. In all the most significant ways, digital preservation of this level most closely mirrors the preservation of physical materials. The digital scholarship most at risk (decaying hardware or software environment, formats approaching obsolescence, etc.) is prioritized above other materials that have a perceived longer life potential.

If preserving the bits and bytes is the default activity for sustaining digital scholarship, the next step is where things get messy. Deciding to “collect”⁶ a piece of digital scholarship goes far beyond just format preservation. Replicating the functionality of the files will largely depend on what one’s integrated support environment can handle. Parameter must be in place to provide the necessary context for collecting since the re-factoring of content may be involved. UVa library partnered with the Institute for Advanced Technology in the Humanities (IATH) in 2000 in the Mellon sponsored Supporting Digital Scholarship (SDS) grant. The goals for this project were to “propose guidelines and document methods for libraries and related technology centers to support the creation and long-term maintenance of digital scholarly projects.”⁷ The original SDS grant forms much of the underpinnings of this current approach. It analyzed digital scholarship from both a technical and a policy perspective. Sustaining digital scholarship can be stated as follows: an attempt to develop a socially and technologically sustainable and scalable model for support and preservation of digital scholarship. The operative words in the statement are *sustainable* and *scalable*. Sustainable gestures to the “trustworthy” nature of the institution (both technologically and conceptually) to continue to support faculty research and scalable to grow those research support models as needed. In order to fully understand the implications of preserving digital scholarship the grant established “levels” of collecting. These break down as follows:

⁵ Some scholars have argued that we need to justify digitizing books based solely on preservation needs. This strategy often leaves the library stuck choosing between preservation and access. Mass digitization is a sound strategy for maintaining access but should only play a part in the overall preservation strategy of an institution. See Hahn’s 2008 article on mass digitization.

⁶ “Collect” in this sense means to migrate the materials into the library managed content environment. Many of these early examples of digital scholarship exist on different servers—not all of which the library manages. Therefore a formal collection strategy needs to be employed.

⁷ SDS Mellon Final Annual Report, 2003.

Level 1: Collecting metadata only – At this level the project would be represented as a single object in the digital library which records that the project exists or existed in the past, and includes some descriptive metadata about the content of the project, people who were associated with it, etc.

Level 2: Saving the project as a set of binary files and metadata only – Only the most basic preservation would be attained at this level. Content files and possibly all the files associated with any custom software would be collected as standard binary files only. The same descriptive metadata would be collected as for level 1, along with technical metadata about the original formats of the files and any software that was necessary to use them. At this level, the assumption is that anyone interested in using the project would be on his or her own in trying to reconstruct it.

Level 3: The content can still be delivered as in the original – At this level, relationships among the content are preserved but no attempt is made to capture the exact action of the project or its look and feel. The user's experience may be different but the ability to navigate the connections that the author provided is preserved.

Level 4: Look and feel intact – The project operates and appears exactly as it was originally intended. The software may not be identical but every effort is made to recreate the user's experience as completely as possible.

Level 5: The project is completely documented – The project is preserved as a complete artifact, documenting its development and history. This could include ephemera such as e-mail archives from a project development team, reviews or citations of the project from other sources, documentation associated with grant proposals, etc.

These levels all map to functionality provided by the integrated repository environment—depending on what level of complexity it can handle. This model is based on the symptomatic reading of the components (derived from a complete technology assessment—see below) and can be adapted to almost any institution's cyberinfrastructure. These can also be thought of as levels of service following recent trends in repository management. William LeFurgy's article "Levels of Service for Digital Repositories" states: "Levels of service can best be thought of as a matrix with one set of values determined by the available technology and the other set determined by the degree to which digital materials have persistent qualities. The first set depends on incremental development of new and improved tools. The second set of values is tied to the degree to which digital materials are persistent (based on

consistent and transparent rules for description and structure, standardized file formats, and so forth)." Embarking on a digital preservation assessment of digital scholarship requires clear guidelines to manage expectations as closely as possible. To outline these activities, it helps to have a formal work plan that can be mapped to a level of collecting.

First order: Do a technical assessment of the digital scholarship. This will also include a census of all the scholarship as defined by the faculty member or as defined by the "collection" or corpus of materials. It is imperative that one undertakes a technical assessment of the scholarship prior to any other activity. This can be broken down into different areas of assessment: technology required, file format, functionality, and intellectual property, digital rights management, to name a few. The assessment should also take into account mappings from current hardware and software environments to the integrated environment that the institution supports. Granted, as with all similar types of activities it can only ever be an approximation but it most certainly can be used to map the project to a level of service (and hopefully, faculty expectations). The first part of any migration (or refactoring as the case may be) is to understand the scope of the scholarship (collection, project) itself. This is a surprisingly difficult process and is often taken for granted that everyone understands the extent of the digital scholarship. In fact, this is seldom the case. This stage is integral to formulating a roadmap of work that will be necessary to digitally preserve the materials for inclusion into the library managed content environment.

Second order: Once the census and assessment is completed you can map the functionality to an appropriate level of service. This should be an agreed upon level between the original manager of the content and the future managers (e.g. faculty member and those responsible for the library managed content environment). If the two parties agree then the next step is to develop and formalize agreements between parties. This could take the form of a collection or deposit agreement and should provide several key components at a minimum:

1. An overview of the intellectual property components of the collection (including copyright and access issues).
2. A formal work plan that maps out each stage of work that will need to be done. This should include shared staff time and server access.
3. Document all decisions and factors related to preserving the digital scholarship so that future managers can understand why certain decisions were made.

Final order: Implement service and procedural methods to formally ingest the digital scholarship into the integrated repository environment. This is also known as the final “publishing” of the digital scholarship. This final stage “freezes” the digital scholarship not allowing any new changes to take place unless governed by the collection agreement. The content is then managed by the library and is digitally preserved to the best of the institution’s ability. This overview is meant to be a conceptual framework that could be adapted to most institution’s missions and infrastructure. It does not do justice to the many complexities and challenges that go into preserving digital scholarship. This process should be mutable and adapted to changing technology and scholars’ needs and is never meant to become a monolithic structure. Digital preservation is still a moving target and we need to be ready to change with it.

Conclusion

When one steps back and surveys the vast complexities involved in the preservation of digital scholarship it becomes painfully clear that unless units across the institution cooperate, we will all fail. The first step is to create a suite of services that can meet our researchers’ needs for supporting and sustaining digital scholarship. Developing a network of cooperative elements to support these services needs to be part of the initial planning. The library, technology units, faculty, provosts, academic departments, all need to have a shared understanding of what the goal for digital preservation should be. The library cannot establish seemingly arbitrary requirements for faculty to manage the scholarly output of the institution, unless the scholars understand what is at stake. University administrators (chancellors, presidents, provosts, deans) all need to agree that the preservation of the scholarly record in a digital world is a complex set of cooperative communication, management, and administration. If the funding is not available for digital preservation then we will fail before we begin. Therefore it is incumbent upon all levels of higher education to understand the implications of a true digital preservation strategy: one that is not bounded by a single department, library, or school; one that is not entirely dependent upon commercial organizations to do it for us; and one that combines all the strengths of librarianship, technology, innovation, and faculty participation. No one can do it alone. Establishing a sound strategy for one’s own institution is only the beginning—partnering with other institutions means that we can begin to develop some digital preservation synergy. We have only just started down this path and there is more to do so that we preserve our scholarly record. Sustaining digital scholarship is the

next phase of approaching collecting faculty output into our cultural heritage. It remains to be seen whether or not we will fully succeed in this endeavor. If we do not, then in Donne’s words, “posterity shall know it too.”

References

ACLS. 2006 “Our Cultural Commonwealth: Report of the American Council of Learned Societies Commission on Cyberinfrastructure for the Humanities and Social Sciences.”

<<http://www.acls.org/cyberinfrastructure/cyber.htm>>.

Entlich, Richard and Ellie Buckley (2006) Digging Up Bits of the Past: Hands-on With Obsolescence, *RLG DigiNews*, Vol. 10, No. 5, 15 October 2006
http://www.rlg.org/en/page.php?Page_ID=20987#article1.

Foster, N. F., & Gibbons, S. (2005). Understanding faculty to improve content recruitment for institutional repositories. *D-Lib Magazine*, 11(1).
<http://www.dlib.org/dlib/january05/foster/01foster.html>.

Hahn, T. B. (2008). Mass digitization: Implications for preserving the scholarly record. *Library Resources & Technical Services*, 52(1), 18-26.

Hitchcock, Steve, Tim Brody, Jessie M.N. Hey and Leslie Carr (2007) Survey of repository preservation policy and activity. *Preserv project*, January 2007
<<http://preserv.eprints.org/papers/survey/survey-results.html>>.

Horrell, J. L. (2008). Converting and preserving the scholarly record: An overview. *Library Resources & Technical Services*, 52(1), 27-32.

Jantz, R., & Giarlo, M. J. (2005). Digital preservation: Architecture and technology for trusted digital repositories. *D-Lib Magazine*, 11(6).
<http://www.dlib.org/dlib/june05/jantz/06jantz.html>.

Lavoie, B., & Dempsey, L. (2004). Thirteen ways of looking at...digital preservation. *D-Lib Magazine*, 10(7/8).
<http://www.dlib.org/dlib/july04/lavoie/07lavoie.html>.

LeFurgy, W. G. (2002). Levels of Service for Digital Repositories. *D-Lib Magazine*, 8(5). Retrieved from
<http://www.dlib.org/dlib/may02/lefurgy/05lefurgy.html>.

Marcum, Deanna B. (2002). “Preservation of Scholarship: The Digital Dilemma” in *The Internet and the University*

Forum, 2002.
<<http://www.educause.edu/forum/ffpiu02w.asp>>

Marshall, C. C. (2008). From writing and analysis to the repository: Taking the scholars' perspective on scholarly archiving. *Proceedings of the 8th ACM/IEEE-CS Joint Conference on Digital Libraries*, 251-260.

Nadal, J. (2007). The human element in digital preservation. *Collection Management*, 32(3/4), 289-303.

Research Libraries Group. (2002). *Trusted digital repositories: Attributes and responsibilities*. An RLG-OCLC Report. Available at:
<<http://www.rlg.org/longterm/repositories.pdf>>

Smith, A. (2007). Valuing preservation. *Library Trends*, 56(1), 4-25.

Adapting Existing Technologies for Digitally Archiving Personal Lives

Digital Forensics, Ancestral Computing, and Evolutionary Perspectives and Tools

Jeremy Leighton John

Department of Western Manuscripts, Directorate of Scholarship and Collections, The British Library
96 Euston Road, LONDON NW1 2DB, United Kingdom
jeremy.john@bl.uk

Abstract

The adoption of existing technologies for digital curation, most especially digital capture, is outlined in the context of personal digital archives and the Digital Manuscripts Project at the British Library. Technologies derived from computer forensics, data conversion and classic computing, and evolutionary computing are considered. The practical imperative of moving information to modern and fresh media as soon as possible is highlighted, as is the need to retain the potential for researchers of the future to experience the original look and feel of personal digital objects. The importance of not relying on any single technology is also emphasised.

Introduction

Archives of 'personal papers' contain letters, notebooks, diaries, draft essays, family photographs and travel cine films; and in 2000 the British Library adopted the term eMANUSCRIPTS (eMSS) for the digital equivalent of these 'personal papers', having begun accepting diverse computer media as part of its manuscript holdings (Summers and John 2001, John 2006).

These media include punched cards, paper tapes, magnetic tapes, program cards, floppy disks of several sizes (8", 5.25", 3.5" and 3"), zip disks, optical disks (eg CDRs and DVDRs) and various hard drives, both internal and external. All three major contemporary operating system families are represented: Microsoft Windows, Apple Macintosh and Unix/Linux as well as earlier systems.

Beyond the library's own collections, the Digital Manuscripts Project has enabled digital capture for the Bodleian Library, the Royal Society (with the National Cataloguing Unit for the Archives of Contemporary Scientists), and the Wellcome Library.

Digital Manuscripts at the British Library

The primary aim of the project is to develop and put into place the means with which to secure the personal archives of individuals in the digital era in order to enable sustained access. This entails the capture of the digital component of the archive alongside its corresponding analogue component.

The project is also addressing in tandem the digitisation of the conventional papers in personal archives (and in that sense is also concerned with digital manuscripts beyond eMSS). Among other benefits, this will make it easier for researchers to work with an entire personal archive in an integrated way; but this work along with cataloguing and resource discovery is beyond the scope of the present paper, which aims to focus on the curatorial role in digital acquisition, examination and metadata extraction.

Theoretical and Practical Considerations

The challenges of technological obsolescence, media degradation and the behaviour of the computer user (eg failure to secure and backup information including passwords) are long familiar to the digital preservation community. Personal collections raise issues, however, that are different from those arising with publications, which have received far more attention.

Of special relevance is the means of acquiring personal archives. Central to the process is the relationship between the curator and the originator or depositor, and in particular the need to deal with personal matters in a sensitive way, ensuring robust confidentiality where necessary.

Three key requirements have been identified and promoted: (i) to capture as far as possible the whole contextual space of the personal computer (the entire hard drive or set of hard drives for example) and not just independent individual files, thereby strengthening authentication; (ii) to replicate and retain exact copies of the original files, recognising their historical and informational value (and not just rely on digital facsimiles, even if these match modern standards for interoperability); and (iii) to meet the special requirements for a confidentiality that is sensitive and reassuring to potential depositors as well as being technically convincing.

A pragmatic philosophy is to provide for immediate access to basic text, images and sounds; but to retain (by capturing and keeping exact digital replicates of disks and files) the potential to make available high fidelity versions that respect original styles, layout and behaviour.

The Digital Capture Imperative

Future work with personal archives can be expected to be increasingly proactive and entail a close understanding with and involvement of originators and their families and

friends. The single most important consequence of the increasingly digital nature of personal archives is the need to preempt inadvertent loss of information by providing advice and assistance.

The key threshold is the initial digital capture: the movement of the eMANUSCRIPT information to modern, fresh and secure media.

Adapting Existing Tools

An effective and potentially efficient route for successful digital capture, preservation and access is to adopt and modify existing technologies for new purposes rather than necessarily designing from scratch.

In this spirit, three key technologies are being examined: (i) computer forensic software and hardware; (ii) ancestral computers, disk and tape drives with associated controllers and software emerging from communities of enthusiasts; and (iii) evolutionary computing techniques and perspectives.

Computer Forensics

In computer forensics there are three text book stipulations (eg Kruse and Heiser 2001, Casey 2002, Carrier 2005, Sammes and Jenkinson 2006): (i) acquire the evidence without altering or damaging the original; (ii) establish and demonstrate that the examined evidence is the same as that which was originally obtained; (iii) analyse the evidence in an accountable and repeatable fashion. There are, moreover, certifiable standards with which computer forensic scientists must comply in order to satisfy legal authorities. Guides to good practice include ACPO (2003) and NIJ (2004). These requirements match in a number of ways the concerns of the digital curator of personal archives.

A wide range of forensic software and hardware has been explored at the British Library for its applicability in capturing, examining and authenticating eMSS. Software that has been and is currently being surveyed and tested includes: Forensic Toolkit (FTK) of AccessData; Macintosh Forensic Suite (MFS) of BlackBag Technologies; Image, PDBlock, DriveSpy and others of Digital Intelligence; Helix of e-fense; Encase of Guidance Software; CD/DVD Inspector of InfinaDyne; Device Seizure, Email Examiner and others of Paraben. Products which have not been examined include ILook Investigator and X-Ways Forensics.

Open source forensic software tools include: Back Track; Coroner's Toolkit; Foremost; Foundstone Forensic Tools; Open Computer Forensics Architecture; Scalpel; and Sleuth Kit with Autopsy.

Forensic hardware includes high specification workstations with forensically compatible BIOS (eg Digital Intelligence), diverse write-blockers (eg Tableau) and robotic floppy disk and optical disk imagers (eg WiebeTech) as well as numerous connectors and adaptors.

Overview of Available Functionality

This equipment provides a plethora of capabilities including the write-protection of original collection source disks, certified wiping of target receiving disk (even brand new drives can contain digital artefacts), the forensically-sound bitstream 'imaging' of the original disk, the creation of unique hash values (MD5, SHA1 and related algorithms) for the entire disk and for individual files, and the recovery of fragments of lost files.

Other functionality includes the ready export of replicate files, the bookmarking and annotating of files of interest for summary reports, timeline viewers for investigating times and dates of file creation, modification and access, while taking into account different time zones, provisional identification of file types based on file signatures and extensions, maintenance of an examination audit trail, filtering of files that are not of immediate interest to an examining curator (eg software files), sophisticated searching (with GREP), file viewing, and reading of emails with carving out of attachments.

Available forensic products are subject to ongoing and rapid development and any attempt to identify the best of them risks being anachronistic. There is no single product that will meet all requirements of the forensic examiner or for that matter the digital curator or preservation expert, which explains why there is a flourishing diversity of specialist products.

Two of the most well established are Encase and FTK, both of which seek to be comprehensive, encompassing in one package much of the functionality just outlined. Both work with a wide range of file systems, and are convenient and comparatively straightforward to use, while still providing capabilities for hexadecimal viewing and analysis of disk and file system geometry. Encase has recently incorporated "Outside In" technology from Stellant for the viewing of files from over 400 file formats. Following its recent major upgrade, FTK now works natively with Oracle's database technology. Other companies such as Paraben provide numerous software modules that are dedicated to specific capabilities and are able to work either separately or together as a more integrated whole with P2 Commander.

On the other hand, CD/DVD Inspector specialises in the analysis of optical discs, which show some profound differences from hard disks in the forensic context (Crowley 2007). A standard ISO 'image' does not capture all of a CD's potential contents, but CD/DVD Inspector is able to do so, producing a file that can be imported into Encase for example. It is also able to work with the sometimes awkward Universal Disk Format.

Helix is another specialist: essentially a forensically customised adaptation of Knoppix. In this Linux mode it serves as a bootable CD with a self-contained operating system that will not write to the attached hard drives, and which can create nonproprietary forensic 'images'. (It also operates in a Windows mode, mainly concerned with the forensics of live machines.) Moreover, it is accompanied by an assortment of other largely standalone tools

(including some of the openly available ones mentioned), making it a kind of forensic Swiss Army knife.

The essential workflow adapted for the curation of information from contemporary computer media in personal digital archives can be considered in two phases: before and after capture.

Phase 1. The Core of the Capture Workflow

There are three initial key requirements: (i) audit trail; (ii) write-protection; and (iii) forensic ‘imaging’, with hash values created for disk and files. (The term ‘disk’ is being used here loosely to refer to floppy, zip, optical and hard disks, flash media and others.)

The first recommended practice is for there to be a chain of custody from the moment that the original materials become available continuing throughout the lifecycle of the entire capture process, recording procedures undertaken by the curator. (At the end of the workflow, the audit culminates with a detailed report.) It is possible to use specialist tools such as Adepto (from e-fense) which will provide an audit log and chain of custody form on acquiring a forensic ‘image’. An advantage of the more comprehensive packages is that the audit control, record making and documentation, is seamlessly integrated and automatic, and in some cases embedded along with the ‘forensic’ image. Digital photos taken by the curator at the time of collection, can be imported into the integrated systems, as can photos of all of the computer media (along with labels) in the personal archive.

The initial motivation for adopting computer forensics arose from the simple concern that even turning on the computer of an originator risked modifying important dates and times of historic interest. It is one of the rules (sometimes needing to be broken) of forensic science not to switch on the originator’s computer (even lifting the lid of some laptops may turn them on); but instead to remove all of the hard drives and connect them to the examiner’s computer using write-blockers.

The main and sometimes necessary alternative to the use of a hardware write-blocker is to again connect the original hard drive to the examining workstation but to boot this computer from a forensically prepared floppy disk or CD, being very careful not to allow the computer to boot from the original hard drive (eg Helix, Encase for DOS, or LinEn).

The long established workhorse of bitstream ‘imaging’ is the ‘dd’ command under UNIX. In principle, this produces a single file encapsulating the entire digital contents of the disk (in practice, it is often a series of conveniently smaller files). An open source forensic version has been developed (dcfldd) with hashes values produced on the fly, and additional features (originally developed by the Department of Defence Computer Forensics Laboratory, and available at sourceforge.net). One drawback of Encase’s compressed ‘image’ file from the perspective of digital curation is its proprietary nature. FTK Imager (which is part of FTK but obtainable separately and free of charge) can create both proprietary and nonproprietary

‘images’ including ‘dd’, as well as computing hash values for disk and files.

The strategy adopted by the Digital Manuscripts Project has been to use both facilities, checking that the same hash values are achieved, as a means of corroborating successful capture, while retaining the nonproprietary ‘image’ file and independently obtained hash values for future reference. It is strongly recommended that digital curators do not rely on any single tool or technology.

Phase 2. Consolidation of the Capture Workflow

The workflow continues with four remaining functional activities: (iv) examination and consideration by curators (and originators), with filtering and searching; (v) export and replication of files; (vi) file conversion for interoperability; and (vii) indexing and metadata extraction and compilation.

With the successful capture of the disk and checks for viruses and other malware completed, examination of its contents can proceed. Sometimes this will be the first time that curator and originator are able to look extensively at the eMSS.

The hash values of the files can be compared with a known hash library for application and operating system software files, allowing these to be identified and filtered out from immediate consideration. Scripts are available and can be customised for refined searching and filtering, based on file signatures, keywords and other criteria. Digital content entailing specific digital rights issues such as intellectual property, data protection or requested confidentiality can be identified by the curator and bookmarked. Any files with credit card numbers, telephone numbers, post codes or email addresses for example, can be automatically located and listed.

Files can be exported in their original form as exact digital replicates providing future scholars with the potential for use with, for example, an authenticated emulator of application software. For more immediate and practical access, the files can be converted into an interoperable form (with low fidelity if not high fidelity) such as a member of the XML or PDF families, where deemed appropriate by the digital preservation community.

Moreover, a digital replicate of the original drive can be restored to a similar or larger hard drive to be inserted into an appropriate computer if desired; however, this presents the same potential problem as before, interacting with this computer will alter the system. The Digital Manuscripts Project is currently examining the use of special and general hardware write-blockers for interacting with a dynamic system.

Encase provides two other options: Physical Disk Emulator (PDE) and Virtual File System (VFS). These modules allow the ‘image’ bitstream to be mounted in read-only mode in a Windows environment. A key difference between them is that PDE, in contrast to VFS, will behave as a normal volume, and not provide access to unallocated space or deleted files. One useful aspect is that these read-only systems can be scanned for viruses and

other malware before exporting any files to the examiner's computer. PDE can be used with the virtualisation software VMware, which will mount the PDE disk as a virtual machine that can be booted virtually. PDE and VFS are both proprietary; but the nonproprietary 'dd' file can also be mounted in VMware (which is itself proprietary though very widely available; open source software such as Xen and QEMU may be useful, however).

InfinaDyne offers a sister tool that can be used to produce a replicate optical disk from the CD/DVD Inspector 'image' file.

The principal forensic tools can conduct deep indexing (incorporating text within files) and extraction of metadata relating to files including file extension, file type, file signature, dates and times, permissions, hash values, logical size, physical location, file extents (fragmentation). In addition, metadata associated with emails (and webmail), photos and instant messages for instance can be extracted. The open source Sleuth Kit (with or without the GUI of Autopsy) is a useful alternative. Metadata for the disks and tapes themselves can be compiled.

Originators, Other Depositors and Third Parties

The essential need to involve potential depositors in the capture process cannot be overemphasised. In addition to assisting in the identification of eMSS where there are data protection and confidentiality requirements (including for third parties), originators can provide contextual and corroborating information that increases the scholarly and historical value of the entire digital archive.

Recovery, even if only in the form of fragments, of partially overwritten, inadvertently or regretfully deleted, earlier drafts of creative works could be of great scholarly interest but it must involve the originators and accord with their wishes. On the other hand, establishing the provenance of fragments of deleted files can sometimes be forensically demanding (Sammes and Jenkinson 2006), and again the creator's confirmation of authenticity might be invaluable. Much better in the long run, of course, would be if creators would know how to manage and care for their personal archive, assisted perhaps by advice from curators and digital preservation specialists.

Passwords are sometimes forgotten or records are accidentally lost, and with the permission of family and originators, decryption and password recovery tools can be used with varying levels of success.

An initial examination of a digital archive can be facilitated at the home of the creator using a forensic laptop and a preview facility that does not entail actual acquisition, helping curators and creators decide whether an archive fits into the collection development policy of the repository before being transferred there. It may be the intention of the originator to simply donate some specific folders or files rather than a disk. A 'logical' acquisition of files can be conducted forensically in much the same way as a 'physical' acquisition of an entire disk.

Ancestral Computing

At present there are, for archival purposes, two limitations to computer forensic technologies: (i) a limited ability to cater for legacy computers, storage media and software even with regard to the initial capture of the information that exists on obsolete media; and (ii) a limited ability to present the files and computer working environment identical or close to the way it was perceived by the creator (even in the case of many contemporary files) with styles, layout and behaviour accurately demonstrated and certified.

It is also necessary to understand the way users interacted with their computers, how these worked technically, the applications that were available to users and the nature of the files produced — just as curators of conventional manuscripts are required to know about the ways in which writing media (wax, parchment, vellum, paper) and associated technologies (pen, ink, pencil, stylus) were designed and used.

This section looks at the initial capture of the information on ancestral computer media. As with deleted files, it is essential to involve originators and their families, as they may not have seen the files residing on the obsolete media for many years.

There is an important and frequently misunderstood distinction between digital capture and digital preservation. Guides to digital preservation have been anxious to dispel any notion of technology preservation as a tenable solution. However, the use of ancestral computer technology for digital capture is unavoidable at present.

Files existing on 3.5" and 5.25" floppy disks and derived from Microsoft DOS and early Windows systems can often be replicated within Windows 98, in DOS mode where necessary, on a relatively recent PC computer furnished with corresponding floppy disk drives. Longstanding forensic tools can help (eg Digital Intelligence's Image, an imaging tool specifically designed for floppy disks).

More challenging are the hundreds if not thousands of species of computer systems which were famously diverse during the 1980s and early 1990s before Microsoft DOS and Windows came to predominate (with varying combinations of processors, operating systems and ROM, and disk systems and actual hardware types) (Nadeau 2002).

Publishing and Typesetting

During and after this period there was a widely felt need to convert files from one type to another, as witnessed by guides to file formats such as Walden (1986, 1987), Kussmann (1990), Swan (1993) and the encyclopaedic Born (1995). The need to create a degree of interoperability in order to move data between applications has long been one of the major motivations for reverse engineering software (eg Davis and Wallace 1997).

One community that required duplication and conversion technologies in the 1980s and 1990s were publishers and

their typesetters, who needed to read and convert files derived from diverse sources (ie writers) to a local standard that could be used by the in-house computer and printing equipment.

InterMedia was a UK company that specialised in supplying media and data conversion systems for over two thousand floppy disk and hardware and operating system combinations. The National Library of Australia has used the system for 5.25" and 3.5" floppy disks (Woodyard 2001).

The company has been bought up by a USA company, eMag Solutions, which retains offices in the UK. An InterMedia system, now renamed eMag Floppy Disk Conversion System Model MMC4000 has been obtained by the British Library with the InterMedia software and Stack-a-Drives for 8", 5.25", 3" and 3.5" floppy disks working with a proprietary floppy disk controller.

One success has been the capture and transfer of files to modern media from 8" floppy disks, dating from a quarter of a century ago. The equipment has also been used to read hundreds of files residing in 3" and 5.25" floppy disks dating from two decades or so. So far there have been relatively few cases where disks have been entirely unreadable: occasionally degradation can be seen in the physical condition of the disk, ie a light reddish brown surface indicative of oxidisation.

Typically the system would have been used to read and convert files, derived from word processors, from one type to another that can be read by modern PCs, using basic Translation Tables as well as program Protocols that can handle pointers. There is a Disk Recogniser function that will sometimes though not invariably assist in identifying disk types. Original files would be converted to a proprietary InterMedia Internal Coding (IMIC), to be subsequently converted to a file with the desired format.

The later version of InterMedia for Microsoft Windows software (IMWIN, Windows XP) is convenient to use but the earlier version for Microsoft DOS (InterMedia, DOS version) is more powerful. It is geared towards a more complete analysis and replication of floppy disks at the most basic levels. Disks can be interrogated at the clock and bit level. In the reading and copying of sectors, disks with hard sectors as well as those with soft sectors can be addressed.

The approach adopted by the Digital Manuscripts Project is, as far as feasible: (i) to copy the individual files in their original format (file digital replicates); (ii) to copy the entire disk (disk digital replicate); and (iii) to create and retain converted files that provide the basic alphanumeric content as low fidelity copies (eg as Word documents) which are later converted to an interoperable form such as PDF.

One simple but useful extension of the overall workflow is to import these files into the forensic system, thereby creating hash values for all the files and integrating them with other files and providing an audit trail.

In addition, MediaMerge for PC (MMPC) has been obtained from eMag Solutions for reading and copying tapes. The user can view and duplicate at block level as well as copying the individual files. A series of 0.25" data cartridges derived from UNIX computers active in the 1990s have been copied by this means.

While it has been very satisfying to capture historically important files using these systems, relying (in the case of floppy disks) on proprietary technology that is no longer fully supported and developed, is clearly not a sustainable solution. The inherent knowledge in this and other data conversion systems is being pursued by several avenues.

Another key source of useful technology for the purposes of the Digital Manuscripts Project has been and will be the classic, retro and vintage computer communities.

Expert Enthusiasts

As a result of continuing enthusiasm for these ancestral computer systems, a small German company called Individual Computers has produced modern technology that enables the reading of early format floppy disks. Specifically, the Catweasel is a universal floppy disk controller that can be used with modern PCs and normal floppy disk drives for 5.25" and 3.5" floppy disks (and in principle others too).

The manufacturer has indicated that Catweasel will work with the following formats (many though not necessarily all variants): Amiga, Apple IIe, early Apple Macintosh, Atari, Commodore and PC, with more planned.

Its attraction lies in its flexibility and degree of openness. Catweasel MK 4 is a low profile PCI card that uses FPGA chips (Field Programmable Gate Arrays) that provide it with reconfigurable logic meaning that software drivers for currently unsupported disk formats can be downloaded when these become available (from Individual Computers itself or expert enthusiasts), and used to reprogram the Catweasel without removing it from the computer. With the appropriate software, it can be used with Linux computers, and use with Mac OS X is anticipated.

The Digital Manuscripts Project has installed the device and is currently exploring its capabilities.

Individual Computers is also involved with other developers in the Commodore One (C-One Reconfigurable Computer). This is a computer that began in 2003 as an enhanced adaptation of the venerable Commodore 64 (C64), one of the most prolific computers of all time. The current version of the C-One (actually a motherboard that can be used with widely available hardware components such as an ATX type computer case) is reconfigurable, again due to FPGA chips. This means that the same basic hardware system can be modified so that it can behave like another early computer such as the C64's sister, the VIC 20, or the Schneider CPC, Atari, or Sinclair Spectrum and others. Expert users are encouraged to create their own FPGA cores using the free development tool Quartus by Altera. Furthermore, with project Clone-A, Individual Computers is developing

a cycle-accurate reproduction of original chipsets in Amiga computers.

Equally this hardware is being matched by software made available for and within the various classic computer communities. Copies of the original software may still be available, as is the case for LocoScript for use with the CP/M operating system running on the Amstrad PCW series of computers.

Other software seeks to emulate the original hardware, operating systems and applications. At the forefront is the Amiga community, for example, with Amiga Forever (preconfigured Amiga ROM, OS and application software files) running on Apple OS X (say) using the UAE emulator of Amiga hardware (Ubiquitous Amiga Emulator). Emulators of application and operating system software are also produced of course which allow early applications to run directly on modern operating systems: for example AppleWin emulates Apple IIe in Windows (available at berlios.org).

There are essentially two sentiments in classic computing: (i) a desire to respect and maintain the original nature of the computer system of interest, down to the exact sounds emitted; and (ii) a desire to ensure the continuing and strengthened relevance of the system by adding modern and new features to it, not least in its interfacing capabilities. This observation and the varying extent to which high fidelity is achieved even when sought points to the crucial role for digital curation and preservation specialists in the certified authentication of these kinds of products. Key institutional resources in this endeavour (in the UK) will be the Science Museum, the History of Computing Museum at Bletchley Park, the Computer Conservation Society, and others, with their expertise and representatives of original equipment.

Along with originators' computers, personal archives frequently contain original software disks and manuals which are likewise retained and used, with permission.

Evolutionary Perspectives and Tools

There are many examples of engineers adapting or copying technologies from nature. Perhaps none is as profound in the digital context as the adoption of DNA itself as a tool.

Digital information, of course, lies at the heart of life in the form of DNA. This has led to the development of DNA computing. But of more direct interest to the present conference is the proposal to use DNA as a means of longterm storage of information (Wong, Wong, and Foote 2003). Three observations have been used to support the idea (Bancroft et al. 2001): (i) viable bacteria have been reported in salt crystals dating from 250 million years; (ii) DNA is the genetic information of humans and therefore will remain central to civilization and scientific progress; and (iii) enormous numbers of identical molecules can be created to ensure informational redundancy to mitigate against stochastic loss.

A vision of DNA encoded library and archival information is a fascinating one but although there are clear advantages to the use of DNA not least in its compact form, the real question to ask is how did it come to be? It is not just a matter of the medium, the molecule, concerned; but the evolutionary process.

Evolutionary Preservation and Capture

Evolutionary science can usefully contribute to digital preservation in a number of ways (John 2006). An aspect of natural selection that is often overlooked is that it reflects the need for diversity in solutions, in strategies. One finds in nature phenomenal amounts of variation; variation that continues to exist generation after generation. It exists because of the inherent unpredictability of nature. It is a recognition — an admission — that the future cannot be predicted. It reflects the existence of multiple strategies: diversity in the face of unpredictability.

It might seem counterintuitive to adopt an evolutionary perspective when striving to preserve something forever (the mission of the British Library's Digital Library Storage system). Quite understandably people tend to marvel at nature's capacity for change but the biological world is also capable of supreme constancy and conservatism. There is information in DNA that has remained the same not merely for thousands or hundreds of thousands of years, but for millions and hundreds of millions of years. It is a phenomenon that deserves the greatest respect of any digital preservation specialist. It confirms the feasibility of deep digital preservation but also points to the need for a humility that seeks more than one best practice, that seeks an evolving strategy incorporating dynamically diverse options.

Conversely, it can be expected that as fundamental advances in digital preservation emerge, it will have many contributions to make to understanding in evolutionary science.

Automation, Information, Personal Archives

Turning to the more directly practical, many powerful techniques of bioinformatics and phylogenetics have been developed where information science meets genome and genetic science. An illuminating example of adapting an existing approach in another field is to be found in the use of phylogenetic algorithms by manuscript scholars wanting to establish or corroborate the ancestry of surviving manuscripts (eg Barbrook et al. 1998, Spencer et al. 2004). These and other bioinformatic techniques will undoubtedly play an important role in authenticating digital files including eMSS; and indeed in forensic analysis of digital files more broadly (eg Goldberg et al. 1998).

There are, however, other aspects of genomic technologies that could be useful in the context of personal archives. The emergence of high throughput gene sequencing capabilities for example has resulted in the production of vast volumes of information, which in turn have led to a demand for automated or supervised computer extraction and interpretation of pertinent information. As a result

dedicated gene and genome databases have been established for remote analysis using GRID and eSCIENCE technologies; but it is not just the protein and gene databases that are available for analysis. There is a burgeoning literature reporting findings from genetic analyses, and in part due to its size this literature too is subject to computational text analysis in its own right (Müller, Kenny, and Sternberg 2004, Raychaudhuri 2006).

Compared with the information in a database, the information in the academic literature (even peer reviewed) is barely structured. Significant advances have been made in the application of natural language processing (Manning and Schütze 1999). One instance is the natural language processing system GENIES which automatically identifies and extracts biomolecular pathways, and forms a key part of GeneWays a technology that processes many thousands of scientific papers and automatically produces a database that is able to identify and visualise molecular relationships and interactions in response to queries from a researcher (Friedman et al. 2001, Krauthammer et al. 2002).

Ontologies will play an important role in testing and training algorithms that provide automated functionality including through supervised machine learning. The expert and ongoing annotation of entities necessary for high quality function coding means, nonetheless, that automation is ultimately going to be necessary to make use of large scale research resources (Raychaudhuri 2006).

The ability to identify the names of genes in scientific literature is not trivial due to inconsistency and nonstandardisation. The most successful algorithms often combine different techniques for classifying documents: descriptive text, nearest neighbour, naive Bayes, maximum entropy and multivariate statistics (Raychaudhuri 2006). The open source software for the multifactor dimensionality reduction technique promoted by the Computational Genetics Laboratory at Dartmouth College, New Hampshire, USA, and used for analysing genes (Moore et al. 2006) has potential for being adapted to pattern search in text.

At first these kinds of technologies will serve less as a means of producing the definitive index, catalogue or ontology, and more as a means of providing pointers, suggestions and indicators for the examiner to confirm independently.

One of the most difficult curatorial challenges of personal digital archives is the need to check for confidentiality and data protection requirements, for copyright issues, for authenticity and provenance concerning all files. Software that was able to automatically search and identify these issues relating to digital rights would be beneficial. It might provide a first stage examination, highlighting likely issues and making suggestions to curators, complementing and strengthening the existing forensic use of GREP searching for example.

It is possible to anticipate in the not too distant future an ability to identify patterns that enable the eMSS to be provisionally classified according to key phases of a

person's life: associated with childhood stages (eg starting school), coming of age, initiation rites, process of a job application, a resignation, a promotion, communications leading to weddings or partnership, professional collaborations, retirement, reminiscence and reflection, births and deaths, memories and remembrance, and so on.

Conclusions

The overall approach of the Digital Manuscripts Project has been in some sense an evolutionary one that allows for flexibility and diversity. It is essential, for example, not to rely on any single technology for digital capture. The adopting and adapting of existing technologies is likewise part and parcel of this approach.

There are a number of existing and evolving technologies that are proving to be useful in the digital curation of eMSS. Software and hardware from the forensic, ancestral computer and bioinformatic communities are evidently useful directly as tools and as sources of ideas and inspiration for digital curators and preservation specialists.

While these existing technologies are providing an urgently needed means of making progress with digital capture, this does not diminish the need for detailed and extensive testing and certification of processes.

Acknowledgements

I am very grateful to Jamie Andrews, Stephen Bury, Katrina Dean, Frances Harris, Scot McKendrick, Ronald Milne, Christiane Ohland, Richard Ranft, John Rhatigan, Elfrida Roberts, Matthew Shaw, John Tuck and Lynn Young for their longstanding and essential support. Arwel Jones, Susan Thomas and Dave Thompson have been reliable and instrumental sources of enjoyable and comradely conversation and discussion.

Although the Digital Manuscripts Project is supported directly by the British Library, the writing of this paper was enriched and made possible by the Digital Lives Research Project, which is funded by the Arts and Humanities Research Council, Grant Number BLRC 8669, and is being led by the British Library. Special thanks to all Digital Lives team members including Andrew Charlesworth, Alison Hill, David Nicholas, Rob Perks, Ian Rowlands and Pete Williams, and most particularly to digital preservation experts Neil Beagrie, Peter Bright, Rory McLeod and Paul Wheatley.

References

- ACPO. 2003. *Good Practice Guide for Computer Based Electronic Evidence*. National Hi-Tech Crime Unit: Association of Chief Police Officers.
- Barbrook, A. C., Howe, C. J., Blake, N., and Robinson, P. 1998. The phylogeny of *The Canterbury Tales*. *Nature* 394: 839-840.

- Bancroft, C., Bowler, T., Bloom, B., and Clelland, C. T. 2001. Long-term storage of information in DNA. *Science* 293: 1763-1765.
- Born, G. 1995. *The File Formats Handbook*. London: International Thomson Computer Press.
- Carrier, B. 2005. *File System Forensic Analysis*. Upper Saddle River, NJ: Addison-Wesley.
- Casey, E. ed. 2002. *Handbook of Computer Crime Investigation. Forensic Tools and Technology*. London: Academic Press.
- Crowley, P. 2007. *CD and DVD Forensics*. Rockland, MA: Syngress Publishing.
- Davis, P., and Wallace, M. 1997. *Windows Undocumented File Formats. Working Inside 16- and 32-Bit Windows*. Lawrence, KS: R & D Books.
- Friedman, C., Kra P., Yu, H., Krauthammer, M., and Rzhetsky, A. 2001. GENIES: A Natural Language Processing System for the Extraction of Molecular Pathways from Journal Articles. *Bioinformatics* 17 Suppl. 1: S74-S82. .
- Goldberg, L. A., Goldberg, P. W., Phillips, C. A., and Sorkin, G. B. 1998. Constructing computer virus phylogenies. *Journal of Algorithms* 26: 188-208.
- John, J. L. 2006. *Because topics often fade: letters, essays, notes, digital manuscripts and other unpublished works*. In *Narrow Roads of Gene Land. Volume 3. Last Words*, ed. M. Ridley, 399-422. Oxford: Oxford University Press.
- Krauthammer, M., Kra, P., Iossifov, I., Gomez, S. M., Hripacsak, G., Hatzivassiloglou, V., Friedman, C., and Rzhetsky, A. 2002. Of truth and pathways: chasing bits of information through myriads of articles. *Bioinformatics* 18 Suppl. 1: S249-S257.
- Kruse, W. G., II, and Heiser, J. G. 2001. *Computer Forensics. Incident Response Essentials*. Boston: Addison-Wesley.
- Kussmann, R. 1990. *PC File Formats & Conversions. Essential Guide to Transferring Data between PC Applications*. Grand Rapids, MI: Abacus.
- Manning, C. M., and Schutze, H. 1999. *Foundations of Statistical Natural Language Processing*. Cambridge, MA: The MIT Press.
- Moore, J. H., Gilbert, J. C., Tsai, C. T., Chiang, F. T., Holden, T., Barney, N., and White, B. C. 2006. A flexible computational framework for detecting, characterizing, and interpreting statistical patterns of epistasis in genetic studies of human disease susceptibility. *Journal of Theoretical Biology* 241: 252-261.
- Müller, H.-M., Kenny, E. E., and Sternberg, P. W. 2004. Textpresso: an ontology-based information retrieval and extraction system for biological literature. *PLoS Biology* 2(11): e309.
- Nadeau, M. 2002. *Collectible Microcomputers. A Schiffer Book for Collectors with Price Guide*. Atglen: Schiffer Publishing.
- NIJ. 2004. *Forensic Examination of Digital Evidence: A Guide for Law Enforcement*. NIJ Special Report. U. S. Department of Justice, Office of Justice Programs, National Institute of Justice.
- Raychaudhuri, S. 2006. *Computational Text Analysis for Functional Genomics and Bioinformatics*. Oxford: Oxford University Press.
- Sammes, T., and Jenkinson, B. 2007. *Forensic Computing. A Practitioner's Guide*. Second Edition. London: Springer-Verlag.
- Spencer, M., Davidsion, E. A., Barbrook, A. C., and Howe, C. J. 2004. Phylogenetics of artificial manuscripts. *Journal of Theoretical Biology* 227: 503-511.
- Summers, A., and John, J. L. 2001. The W. D. Hamilton Archive at the British Library. *Ethology, Ecology & Evolution* 13: 373-384.
- Swan, T. 1993. *Inside Windows File Formats*. Reading, MA: Addison-Wesley.
- Walden, J. 1986. *File Formats for Popular PC Software. A Programmer's Reference*. New York: John Wiley & Sons.
- Walden, J. 1987. *More File Formats for Popular PC Software. A Programmer's Reference*. New York: John Wiley & Sons.
- Wong, P. C., Wong, K. K., and Foote, H. 2003. Organic data memory using the DNA approach. *Communications of the ACM* 46: 95-98.
- Woodyard, D. 2001. *Data recovery and providing access to digital manuscripts*. Information Online Conference 2001, Sydney.

Enduring Access to Digitized Books: Organizational and Technical Framework

Oya Y. Rieger

Cornell University Library
Ithaca, NY, USA
oyr1@cornell.edu

Bill Kehoe

Cornell University Library
Ithaca, NY, USA
wrk1@cornell.edu

Abstract

The digitization of millions of books under corporate and non-profit programs is dramatically expanding our ability to search, discover, and retrieve published materials. Accompanying this progress are cultural heritage institutions' concerns about the long-term management challenges associated with providing enduring access to a large corpus of digitized materials, especially within the confinements of copyright laws. The goal of this presentation is to describe Cornell University Library's program to illustrate a range of organizational and technical issues involved in planning and implementing a preservation infrastructure for digitized books.

Large-scale digitization of published materials has brought millions of books hidden in library stacks to the public eye, making them easy to identify and locate. During 2006-2007, when Cornell University Library (CUL) signed contracts with Microsoft and Google to embark on two large-scale digitization initiatives, the Library staff was equally excited and anxious about the new roles and responsibilities required to successfully manage such a program.

The Library has been involved in various digitization initiatives since the early 1990s; however, given limited funding and the available digitization technologies, CUL had managed to digitize only close to 12,000 books by 2006. At this rate, it would have taken us hundreds of years to convert our entire collection of 7 million items. Whereas the Microsoft collaboration, which lasted for 18 months, resulted in the digitization of close to 100,000 public domain books.

The Google digitization collaboration, which is still in the initial planning stages, involves digitizing approximately 120,000 books per year for five years, covering both public domain and in-copyright materials. In addition, although at a significantly lower pace, there is an in-house digitization operation that grew out of the Microsoft collaboration to systematically digitize special and rare

materials from the Library's collection. The goal of this article is to describe the preservation infrastructure under development that will ensure the effective management of these digital assets.

Preservation Framework

The Cornell University Library drafted its first digital preservation policy framework in 2004, formalizing the library administration's ongoing commitment to the long-term preservation of its diverse digital assets. Although a strong mandate was articulated and the policy included a range of operating principles, roles, and responsibilities, the policy did not move into an implementation stage until the launching of the large-scale digitization initiatives. The prospect of assuming the responsibility of a large body of digital content prompted the library staff to take quick steps to develop a preservation program.

The three legs of the Cornell digital preservation program include *organizational framework*, *technological infrastructure*, and *resource requirements*. Utilizing this three-tiered approach, the following sections describe the decision-making and implementation processes for CUL's preservation program for digitized books. The original three-tiered approach has been expanded to incorporate *access mandate*, which has a critical value for current and future scholarship.

Organizational Framework and Policy

Throughout the last 15 years, we have learned from first-hand experience that technologies alone cannot solve preservation problems. Institutional culture, policies, strategies, staff skills, and funding models are equally important. Organizational infrastructure includes policies, procedures, practices, people – the elements that any programmatic area needs to thrive, but specialized to address digital preservation requirements.

Digital preservation requires a sequence of decisions and actions that begin early in the life cycle of an information object. Standard policies and operating principles for digital content creation are the foundation of a successful preservation program. The critical components include:

- Technical specifications for content creation to specify image-quality parameters for archival and derivative files;
- Requisite preservation metadata with descriptive, administrative, structural, and technical information to enhance access, enable content management, and facilitate discovery and interoperability;
- Quality control and assurance protocols for digital images and associated data.

Although the Library had established digitization and metadata standards prior to the initiation of the large-scale conversion project, we had to reassess our requirements within the scope of our collaborations with Microsoft and Google. Due to the collaborative nature of the initiatives, the companies' digitization protocols and target outcomes set the parameters for digital content creation process.

As the Library was negotiating the contracts with Microsoft and Google, the University Librarian appointed a team called Large-Scale Digitization Steering Committee to oversee various phases of the initiatives with a holistic approach, from selection and preparation of materials to ingest and archiving of digital books. In addition, the Committee was charged with the critical process of identifying staff skills and patterns (and associated costs) required to implement digitization and preservation strategies. One of the Committee's first challenges was to define a new set of requirements that could be supported by the technical provisions of the corporate partners – to compromise between what was available with what was desirable. Some of these technical decisions are illustrated in the following section.

An example from the Committee's current agenda involves exploring our legal rights to preserve in-copyright content. Although the Library's Microsoft project focused on public-domain materials, the collaboration with Google includes 500,000 books representing both in- and out-of-copyright materials. We have a myriad of question to address. For example, is it legally permissible for a library to rescan originals that are not in the public domain to replace unusable or corrupted digital objects? What are the copyright implications of migrating digital versions of materials in copyright from the TIFF to JPEG2000 file format? Section 108 of the U.S. Copyright Law articulates the rights to and limitations on reproduction by libraries and

archives; however, the right to take action to preserve digitized content that is copyright protected is still under study by the Section 108 Study Group convened by the Library of Congress.

Technological Infrastructure

E-science data initiatives have introduced libraries to the challenges associated with large-scale database storage and retrieval. Nonetheless, many participating libraries still have limited experience in data management at the scale of these initiatives, even though the technology that makes preservation possible has the same basic components as the technology of digital collections. The following sections highlight some of the important components of our technological infrastructure, especially from decision-making perspectives.

JPEG2000 as an Archival File Format

The page image files in our digital archive constitute 97 percent of the space required to store the digital books. The format used for storing the images has become important not only from the perspective of best practice for digital preservation, but also from the economic view of sustainability over the long term. Fortunately, best practice and fiscal prudence meet in the JPEG2000 format. Others have reported on the archival benefits of the format—for example, its capacity to embed metadata and yield scaled derivatives easily. Lastly, its ability to be compressed without significant visual degradation translates into significantly lower storage costs.

Physical Storage

For most of its servers, the Library contracts with Cornell's central information technologies group for maintenance and storage. That arrangement proved most cost-effective when we investigated the options for large-scale storage. At the beginning of our search, we expected to store JPEG page images and assumed a need for about 100 terabytes. Our decision to convert the JPEGs to the JPEG 2000 format reduced our storage need by more than 60 percent, and a 40-terabyte array of 1-terabyte SATA drives from Digi-Data Corporation satisfied our requirements for a unit of storage. One unit was sufficient for the first year of production (although we expect to make additional unit purchases in the coming years). The disks are being managed on a three-year lifecycle as a write-once array, in order to minimize maintenance. Deletions are discouraged—a maintenance policy that is easily met by our preservation policy, which demands that nothing be deleted and that any updated objects are added as new versions of earlier objects.

Redundancy Arrangements

Backing up terabytes of data to tape, even static terabytes that aren't expected to change, is a slow, cumbersome process. Restoring a large-scale system from tape would also be very slow. The Library has chosen to assure redundancy by keeping copies of the archived objects on remote storage arrays. Partners with access to Internet2 can speed copies to us if necessary. To mitigate the risk of losing our metadata, however, the XML containers are being backed up to tape locally.

The Choice of an Archival Storage System

After having decided that we would not build a data management and archival storage application ourselves, we examined the characteristics of aDORe and Fedora. We set up test implementations of each and experimented informally with ingest and access. Both systems showed themselves to be capable of managing complex objects

well. At the time we investigated the systems, Fedora was the more flexibly access-oriented of the two, while aDORe had the more stable indexing mechanism for an object's component files. Even though Fedora's large user community and its flexible object model were very attractive, aDORe's storage model—its use of the Internet Archive's ARC-file format and cross-indexed XML metadata containers—promised to use our storage array more efficiently. With our primary focus on the archiving our digitized books rather than providing public access to them, we chose to base our system on aDORe. Nevertheless, we appreciate Fedora's capabilities and plan to use it as the middleware framework for a user-oriented access system as well as reassessing our decision to use aDORe.

Illustration 1: High-level view of the aDORe Archive system

(from <http://african.lanl.gov/aDORe/projects/adoreArchive/>; used by permission of Los Alamos National Laboratory Research Library)

Archival Storage Architecture

The Los Alamos National Library's aDORe archive is a self-contained archival storage system based on the OAIS Reference model. The core is a dual-format storage mechanism: Metadata about complex objects is aggregated in a format called XMLTape; the datastreams that constitute the objects' files are stored in the ARC file format originated at the Internet Archive. The OpenURL's pointing to the datastreams are indexed for ease of

retrieval. References to the datastreams are embedded in the XMLTapes. An index of identifiers and timestamps enables OAI-PMH access to the data through the XMLTapes.

Objects to be ingested must first be described in an XML format; Cornell uses a METS container. An external database is used to provide mapping between Descriptive Metadata and aDORe OpenURLs for administrative and user access.

Metadata Requirements

Preservation metadata incorporates a number of emphasizes recording digital provenance (the history of an object). Documenting the attributes of digitized materials in a consistent way makes it possible to identify the provenance of an item as well as the terms and conditions that govern its distribution and use.

The role of technical metadata (or lack thereof) in facilitating preservation activities is not yet well documented. Although incorporated in preservation metadata, technical metadata merits special mention because of its role in supporting preservation actions. Published in 2006, ANSI/NISO Z39.87 Technical Metadata for Still Images lays out a set of metadata elements to facilitate interoperability among systems, services, and software as well as to support continuing access to and long-term management of digital image collections. It includes information about basic image parameters, image quality, and the history of change in document processes applied to image data over the life cycle. The strength and weakness of Z39.87 is its comprehensive nature. Although in many ways an ideal framework, it is also complex and expensive to implement, especially at image level. While most of the technical metadata can be extracted from the image file itself, some data elements relating to image production are not inherent in the file and need to be added to the preservation metadata record.

It is difficult to consider an image to be of high quality unless there is requisite metadata to support identification, access, discovery, and management of digital objects. Descriptive metadata ensures that users can easily locate, retrieve, and authenticate collections. CUL relies on bibliographic records extracted from local Online Public Access Catalogs (OPAC) for descriptive metadata. Compared with early digitization initiatives, minimal structural metadata are captured. We are committed to use of a persistent IDs to ensure that globally unique IDs are assigned to digitized books; however, we have not yet developed an access system to address this requirement. We do not capture detailed structural metadata, which facilitates navigation and presentation by providing information about the internal structure of resources, including page, section, chapter numbering, indexes, and table of contents.

Resource Requirements: Understanding Financial Implications

Some digitization costs such as materials shipping, scanning, processing, OCR creation, and indexing are covered by Microsoft and Google. However, staff members at the Library are supporting these initiatives by spending significant amounts of time negotiating,

categories, including descriptive, administrative and structural. PREMIS metadata planning, overseeing, selecting, creating pick lists, extracting bibliographic data, pulling and re-shelving books, and receiving and managing digital content. This is an exhausting and disruptive workflow, and its associated local expenses are significant.

During Fiscal Year 2008, Cornell University Library invested close to seven full-time equivalent staff (distributed among a total of 25 staff members) in managing LSDI-related tasks for digitizing 10,000 books a month. It is difficult to calculate a fixed cost because of individual factors that affect selection and material-preparation workflows and the varied physical environments at participating institutions. Different staffing configurations are also required for ramp-up versus ongoing processes. Often neglected or underestimated in cost analysis are the accumulated investments that libraries have made in selecting, purchasing, housing, and preserving their collections.

Although our initial preservation strategy is comprehensive and treats all the digitized books equally, one of the questions we need to explore is whether we should commit to preserve all the digital materials equally, or implement a selection process to identify what *needs* to be preserved, or assign levels of archival efforts that match use level. According to a widely cited statistic, 20 percent of a collection accounts for 80 percent of its circulation. An analysis of circulation records for materials chosen for Cornell University Library's Microsoft initiative showed that 78 percent to 90 percent of those items had not circulated in the last 17 years. In Cornell's case, the circulation frequency may be lower than average because of the age of the materials sampled: all were published before 1923.

Because selection for preservation can be time-consuming and expensive, the trend will likely be to preserve everything for "just-in-case" use. The long-tail principle also may prove that every book finds its own user when it is digitized and discoverable on the Web.

Access Mandate

The 800-pound gorilla in the Library's preservation agenda is the future of Web access to digitized books. Several staff members expressed concerns that digital content may no longer be available in the future through present-day search engine portals, which evolve rapidly in terms of both content and retrieval technologies.

The May 2008 announcement about the closure of the Microsoft Live Search Program proved that the apprehension was not unwarranted. The Microsoft Live Book search website was closed down as soon as the

announcement. Because the Library was relying on using the Persistent IDs provided by Microsoft to connect users from its online catalog to digital books, the unexpected development caused a reroute to square one in means of exploring access options.

Currently, the Library has plans in place to implement bit preservation. However, providing enduring access by enabling online discovery and retrieval of materials (within limitations of copyright laws) for future generations is an enormous challenge—one that may not be met unless faced collectively by research libraries. Efforts at the individual library level will not adequately address the enduring-access challenge unless there is a plan for providing aggregated or federated access to digital content.

From scholarship perspective, the scale of the digitization undertakings is exhilarating and introduces the possibility of novel ways of finding and analyzing content that have been historically presented in print formats. Today's users prefer searching and retrieving information in integrated search frameworks and use digitized books only if they are conveniently accessed at their preferred search environments and support their searching and reading preferences. Therefore, hosting public domain digitized books solely through individual library portals is likely to be insufficient. Having more than one search engine host the same content is likely to increase the survival of digital materials.

Although today's users typically prefer to search for resources online, recent surveys and anecdotal evidence suggest that many users continue to favor a print version for reading and studying—especially for longer materials such as books. This is especially true for humanists as their scholarship heavily relies on close reading and interpretation of texts. CUL has been using the print-on-demand service provided by Amazon/BookSurge to make digital content created through institutional efforts available for online ordering. Thus far the initiative has been limited to the books digitized through past digitization initiatives. The Library is in the process of assessing the PoD options for public domain materials digitized through Microsoft collaboration.

Concluding Remarks

Large-scale digitization initiatives have been unexpected and disruptive—at least for some of the participating libraries such as Cornell. The initiatives began at a time when we are actively exploring our programs in light of developments such as Google's search engine for information discovery and a growing focus on

cyberinfrastructure and the systems that support data-intensive initiatives. There is also increasing pressure to focus digital preservation efforts on the unpublished and born-digital information domain, where preservation concerns are most urgent.

Although research and practice indicate that users increasingly prefer digital information and services, academic and research libraries remain under pressure to continue traditional services too. It is rare to hear about a service being eliminated in order to shift funds into a newly growing area. But the costs of processing and archiving new digital material may cause a significant shift in how funds are distributed among services at many libraries. It is important to try to articulate a preservation program for digital books within the broader scope of library activities and mid-term strategies. Also critical is to envision digital preservation and enduring access by taking into consideration evolving scholarly needs and various information genres and formats.

Acknowledgments

We would like to thank the members of the LSDI team in planning and implementing the digitization project at Cornell. The technical aspects of the project have greatly benefited from vision and hard work of Danielle Mericle, Adam Smith, Marty Kurth, Jon Corson-Rikert, John Ferreira, and Frances Webb.

References

The following report provides an overview of challenges faced in large-scale digitization of library materials: Oya Y. Rieger. *Preservation in the Age of Large-Scale Digitization*. Washington, DC: Council on Library and Information Resources, 2008, <http://www.clir.org/pubs/abstract/pub141abst.html>

Cornell University Library, Digital Preservation Policy Framework, 2006. <http://hdl.handle.net/1813/11230>

Anne R. Kenney and Nancy Y. McGovern, "The Five Organizational Stages of Digital Preservation," in *Digital Libraries: A Vision for the Twenty First Century*, a festschrift to honor Wendy Lougee, 2003.

Section 108 Study Group Report, March 2007, <http://www.section108.gov/docs/Sec108StudyGroupReport.pdf>

Buonora, P. and Liberati, F. (2008). A Format for Digital Preservation of Images: A Study on JPEG 2000 File

Robustness. D-Lib Magazine, 14(7/8)
<http://www.dlib.org/dlib/july08/buonora/07buonora.html>

aDORe: <http://african.lanl.gov/aDORe>

Fedora Commons: <http://www.fedora-commons.org>

PREMIS: <http://www.loc.gov/standards/premis>.

Z39.87: Data Dictionary—Technical Metadata for Digital Still Images. Available at
http://www.niso.org/standards/standard_detail.cfm?std_id=731.

Metadata-extraction tools such as JHOVE and NLNZ Metadata Extractor Tool generate standardized metadata that is compliant with PREMIS and Z39.87.

Lorcan Dempsey. 2006. "Libraries and the Long Tail: Some Thoughts about Libraries in a Network Age." *D-Lib Magazine* 12(4). Available at
<http://www.dlib.org/dlib/april06/dempsey/04dempsey.html>.

Book Search Winding Down, May 23, 2008,
<http://blogs.msdn.com/livesearch/archive/2008/05/23/book-search-winding-down.aspx>

According to a study at the University of Denver, most of the problems people perceive with electronic books are related to the difficulty of reading large amounts of text on the screen. Michael Levine-Clark. 2006. "Electronic Book Usage: A Survey at the University of Denver." *Libraries and the Academy* 6(3): 285-299.

Creating Virtual CD-ROM Collections

Kam Woods*, Geoffrey Brown**

*Indiana University
Department of Computer Science
150 S. Woodlawn Ave., Bloomington, IN 47405-7104
kamwoods@cs.indiana.edu

**Indiana University
Department of Computer Science
150 S. Woodlawn Ave., Bloomington, IN 47405-7104
geobrown@cs.indiana.edu

Abstract

Over the past 20 years, more than 100,000 CD-ROM titles have been published including thousands of collections of government documents and data. CD-ROMs present preservation challenges at the bit level and in ensuring usability of the preserved artifact. We present techniques we have developed to archive and enable user access to a collection of approximately 2,900 CD-ROMs published under the Federal Depository Library Program (FDLP) by the United States Government Printing Office (GPO). The project provides web-based access to CD-ROM contents using both migration and emulation and supports remote execution of the raw CD-ROM images. Our project incorporates off-the-shelf, primarily open-source software. The raw data and (METS) metadata are made available through AFS, a standard distributed file system, to encourage sharing among libraries.

Introduction

CD-ROMs present significant preservation challenges. At the bit level, the obvious technique is to create an (ISO) image of the standard ISO-9960 file system; however, CD-ROMs are subject to bit rot and generally do not provide checksum information to determine if an image is error-free. Building a viable archive requires comparing images from multiple instances of a single item. This can be viewed as an inverse of the problem solved by LOCKSS – lots of copies are required to create a single reference image which might itself be preserved through a system such as LOCKSS (Maniatis et. al. 2005).

At the usability level, ISO images are large (up to Gigabytes), must be “mounted” to enable access to their contents, and often require software installation in order to use those contents. Furthermore, the required software is quickly becoming obsolete. We assume a use model in which most patrons will be satisfied with the ability to browse within a CD-ROM and easily access documentation and data in obsolete formats. For a minority, access requires mounting and executing software from the CD-ROM on a physical or virtual machine. Many items, although not the FDLP materials, require authentication in order to ensure that copyright restrictions are satisfied.

Our research has focused on handling the FDLP document collection held by the Indiana University Libraries. This collection represents an ideal research

workload because it is an important preservation target, it is temporally and technologically diverse, and it presents few copyright restrictions. The techniques we describe generalize to other CD-ROM based materials. Searching the Indiana University library reveals more than 14,000 items. A similar search of the OCLC Worldcat system reveals more than 120,000 items.

The GPO has published approximately 5000 unique CD-ROMs and DVDs created by various government agencies and distributed these publications to various subsets of the 1450 depository libraries (CD/DVD Database 2007). These collections contain fundamental information about the economy, environment, health, laws and regulations, and the physical and life sciences. The technological span of the collection ranges from items created for MS-DOS and Windows 3.1 – requiring execution of proprietary binaries – to recent items relying exclusively on commonly available commercial applications.

The FDLP is organized as a hierarchy of state-level regional repositories holding complete collections and all other repositories holding subsets (FDLP 2006). A patron wishing to access a particular item must first locate a repository holding the item and then obtain physical access to typically non-circulating materials. The libraries support physical access to CD-ROMs through “reference” workstations mandated by the GPO (MTR 2005).

Our project is creating a virtual collection of CD-ROMs accessible from any Internet enabled location. The collection is browsed via a web-server and the CD-ROM images accessible through a distributed file system (AFS). In a typical use-case, a patron searches the database to find items of interest, browses those items to determine suitability, and mounts images on a physical or virtual workstation. It is anticipated that libraries will utilize standard virtual machine (VM) technologies to replace existing reference workstations. Our project includes script development simplifying the use of a VM to access the collection.

In contrast with traditional repository models, our objective is to enable libraries to integrate a collectively maintained “virtual collection” into existing collections. By retaining the images and metadata in AFS, libraries are empowered to pool otherwise disparate resources. The web-browsing capabilities can be seamlessly integrated into a library’s infrastructure, while AFS

provides the means to maintain and adjust access privileges over multiple Kerberos domains.

The remainder of this paper focuses upon the techniques and tools used to build a web-based project providing browsing and execution of CD-ROM collections. The technical discussion is divided into two sections. The first deals with accessing CD-ROM contents including file access, format identification, web-based browsing, migration, and the use of virtualization tools to support legacy executables within CD-ROM images. The second section includes CD-ROM image preservation and distributed image access. Throughout this paper we refer to ISO images which are the “bit faithful” copies of CD-ROMs; ISO is short for ISO9660, the standard data format for CD-ROM contents (ECMA 1987). We conclude with a discussion of related work.

File Access

Given a collection of ISO images of CD-ROMs, and the ability to read the files contained within these CD-ROMs, how can we ensure the continued utility for these files? As we discuss in the Image Access section, preserving raw access to these files is “easy”. However, ensuring their continued utility in the face of obsolescence is hard.

The foundation for our experimental work is a basic web service that supports search, browsing, and migration to modern formats; however, this web service is intended purely as a demonstration vehicle. Our overall approach utilizes open-source software libraries and tools that can be integrated into existing collections. Indeed, the core web service is simple – requiring approximately 750 lines of Perl. However, in building this service we were forced to address fundamental issues involving file access (discussed further in the Image Access section), file format identification, browsing of files with hardwired context dependencies (e.g. HTML), and building reliable migration services. Because many of the FDLP items are dependent on proprietary or obsolete binaries, we assume that committed users will need to utilize emulation (virtualized) execution environments. A part of our research effort has explored the use of automation to simplify emulation based access, including the creation of automated installers for proprietary applications.

The remainder of this section is organized as follows. We begin with an overview of our web-service to introduce the fundamental issues, including file browsing in the face of contextual dependencies. We then consider format identification, file migration, and finally emulation.

Web Service

The web service we developed to access the FDLP collections has a conventional user interface – a user finds items of interest through a search screen; these items are presented in a manner similar to a conventional library catalog. The metadata listing for each item

provides links enabling the browsing of ISO image contents or access to the raw image. Browsing within an image is analogous to a file browser with file title, type, size, and creation date. Individual files may be accessed in original format or migrated rendition.

The web service is driven from AFS accessible ISO images and corresponding metadata in METS format. The search indices and human readable “catalog” pages are generated from the METS metadata through XSLT transformation. Browsing within ISO images is supported by separate binaries to identify formats, extract files and directories, and migrate files to modern renditions. This partitioning is intended to make integration of the underlying technologies into existing library collections “easy” – the web service is a relatively thin code veneer binding the raw data and metadata with tools supporting browsing.

An important design decision is that all ISO images and their constituent files appear to reside within a static file system hierarchy. In an earlier implementation, which was indexed by Google, we found it difficult to locate the context of an individual file. In our current implementation, the context of a file can be found by “walking” up the URL from file to enclosing directory to image, and ultimately to the catalog metadata describing an item. URLs for migrated renditions are encoded as HTML “gets” based on the original URL. For example, an Adobe PDF rendition of `../foo.doc` is accessed using `../foo.doc?migrate=pdf`. The listings for directories containing files which have migrated renditions provide appropriate icons for accessing those files. As illustrated above, the original rendition of a migrated file is easily located by dropping “? ...” from the corresponding URL.

A significant problem with browsing arises from links. For example, an HTML file within some ISO image may refer to pictures or other HTML files. These links may be relative (e.g. `foo/bar.html`) or absolute (e.g. `/foo/bar.html`). Unfortunately, the latter implicitly refers to the root of the ISO image rather than the root of the web-server. In our system we found it necessary to interpret and patch HTML files as they are served in order to ensure that the browsing experience works as expected. Unfortunately, such patches are not always feasible. For example, PDF files may have embedded links, or HTML files may use links within javascript. Hence, our patching is good, but imperfect.

We use Swish-e (Simple Web Indexing for Humans - Enhanced) to provide indexed search via the “title”, “abstract”, “subject”, and “classification” categories drawn from the METS records associated with each ISO (Rabinowitz 2004). A default query made in the web interface is searched by title only, although more sophisticated searches using boolean operators, wildcards, or requesting specific SUDOC number(s) are also handled.

Each query returns a number of hits corresponding to catalog records. Selecting an individual hit returns a page providing the full formatted METS record, along with links to browse the contents of an ISO or download the

full image. Information about current directory location and file format of any object within that directory is provided to Apache by a series of Perl CGI scripts processing the ISO using the `libiso9660` library and the Freedesktop Shared MIME database (Leonard 2008). Additional information about this method is provided in the Image Access section.

A user browsing the ISO is presented with a modified Apache-style listing for file objects within the directory hierarchy. For each object this includes an icon selected according to MIME type, a link to a migrated rendition (if available), name, modification date, and size. Formats for migrated renditions (including - primarily - HTML and PDF) are chosen for ease of access within a standard browser.

Figure 1: Web server overview.

The web service described here is modular, designed to be used for standalone access to independent CD-ROM collections or integrated into existing archival systems. Migration services are therefore loosely coupled to the rest of the service, and may be run on a dedicated server. A simplified representation of the web service backend, along with a typical client setup as discussed in the Emulation section, is given in Figure 1.

Object Format Identification

Accurate file format identification is critical both to the presentation of sensibly marked document links by the web service and to the automated server-side migration services. Identifying specific files for which migrations can be performed is particularly difficult, since any file format identification scheme will generate a certain percentage of false positive hits for each document type. Our strategy uses a combination of existing open-source tools for identification along with additional scripted tests and heuristics to provide breadth of coverage while tolerating failures gracefully.

We use the open source Shared MIME-info Database specification developed by the X Desktop Group for primary identification. In particular, we use the `libsharedmime` implementation found in the current

distribution of the Gnome desktop. This has a number of advantages over other available file format registries (PRONOM 2008), (GDFR 2008). It is production quality, fast, integrated into the Unix environment we use for migration services, has an easily customizable database, and produces succinct machine-readable descriptions of identifications. It remains in active development, and specialized (complementary) database updates for field-specific (for example, chemistry or GIS) file types are readily available.

File extensions and simple analysis for binary content are used as secondary identifying characteristics. This provides a degree of flexibility in handling the original file object. As an example, trials on the CD-ROM collection have indicated that both the Shared MIME-info Database and preservation-specific tools such as DROID will generate false positives or tentative hits for documents with the `.doc` extension containing some binary data that are not, in fact, Microsoft Word documents - or in certain cases cannot be migrated to a modern format using the OpenOffice document filters without damage or data loss. The secondary identifying characteristics allow for a more finely-grained distinction between conversion failures and may generate fallback conversions when required (for example, text-only extraction for those office documents where binary content is mangled or cannot be appropriately identified). These generated materials are intended primarily to improve collection access rather than address the multitude of technical and preservation issues with long-term format migration.

Migration

Our project tracks a diverse set of candidate file types for format migration. The web interface streamlines access by providing links to migrated renditions of original materials. Examples include Microsoft Office documents, Lotus 1-2-3 files, media items, and scientific binary formats. On user request, these migrated renditions are generated unless previously cached. As previously discussed, contextual information is used rewrite HTML sources where necessary for browsing, such as for archived websites with broken absolute site-internal links.

We use a collection of open source migration tools along with a control scripts to create migrated renditions of documents in legacy formats. Our emphasis is on leveraging existing frameworks - the Shared MIME-info database, OpenOffice format filters, and the Python-UNO OpenOffice API bridge - and server-side scripting to provide both on-demand and batch migration paths for each selected format. Our approach tolerates and logs conversion failures in the background. The web service provides links to exactly those files for which successful conversions have been performed.

A Python framework coordinates both batch and on-demand conversion tasks. For batch conversions, a high-performance subprocess is called to rapidly generate a walk of the content within one or more ISO images. The

results are filtered for the requested conversion formats, and written to a separate log for each ISO.

Converted documents are stored on a dedicated AFS volume. Each converted document is uniquely renamed using an MD5 hash constructed from the absolute ISO-internal path name, avoiding collisions and allowing for a simple single-directory storage path for each ISO in the collection.

Microsoft Word and PowerPoint documents are converted to Adobe PDF via a daemonized server which allocates “headless” instances of OpenOffice 2.4 to each conversion task. Conversion failures are automatically logged by the server. Additionally, the server monitors the health of each OpenOffice instance. If memory usage exceeds an administrator-defined level, or if a conversion task appears to be hung (over time measured on a sliding scale according to the size of the document), the instance is killed and the failure logged, maintaining system stability.

A similar mechanism monitors the conversion of Lotus 1-2-3 and Microsoft Excel documents to browser-friendly HTML or structured XML using Gnumeric 1.8.3 (current stable release as of writing) and appropriate filters. Microsoft Access and DBase III/III+/IV files are converted using Python modules to extract data directly from the known binary format. Additional media documents, including MPEG video and DVD video files, are migrated to flash video in a web-friendly resolution.

In our dispatch model, the AFS client provides access to and retrieval of materials held on an off-site AFS server (as shown in Figure 1). Migration requests consist of a sequence of one or more files - or a top-level directory containing the objects to be migrated - along with target format(s) and the known communications port for the OpenOffice daemon. The daemon itself is multi-threaded, and communicates with each migration instance on a unique port. The Python code may be customized to a specific site installation via a simple XML configuration file. It is based on an OpenOffice daemon script provided by OpenOffice.org along with the GPL-licensed daemon distributed as part of the ERP5 Enterprise Resource Planning system (ERP5 2008).

In previous work, we collected statistics on the distribution of file formats within the FDLP collection at the Indiana University Libraries, and provided results demonstrating the feasibility of automated migration from legacy formats (Woods and Brown 2008). While we maintain the ability to migrate all previously examined file types, our trials in this work focused on stress testing the automated migration environment. Using the format identification procedures discussed earlier, we selected 41403 Microsoft Word documents, 23569 Microsoft Excel documents, and 24780 Lotus 1-2-3 documents for batch conversion. The server successfully managed each migration task, logging conversion failures and terminating frozen instances of OpenOffice as required. The migrated materials are maintained alongside the source ISO images on the AFS.

Emulation

The current model for utilizing the FDLP materials requires physically mounting a CD-ROM on a “reference workstation.” The GPO requires depository libraries to maintain such workstations for patron use and provides specifications for the required software. ISO images can similarly be mounted utilizing common tools such as “daemon tools” and hence require no fundamental change in utilization model; however, there are significant problems with the use of reference workstations that can be ameliorated through virtualization. Furthermore, the transition to a virtual CD-ROM collection offers the opportunity to eradicate the geographical barriers implicit in the current model.

While the GPO provides specifications for reference workstations (<http://www.fdlp.gov/computers/rs.html>), these are updated periodically to reflect new requirements without ensuring continued access to older items. Indeed, the GPO (<http://www.fdlp.gov/computers/rsissue.s.html>) states that:

Libraries should also consider keeping [existing] equipment in order to access electronic products that cannot be read with newer hardware and software.

For sparsely used materials, this suggestion seems problematic. Furthermore, many of the GPO items require installation procedures that mutate the software environment in potentially incompatible ways. A natural solution to both the problem of maintaining older reference workstations and clean environments is emulation (virtualization) which requires only that hard disk images (including operating system and applications) be preserved. Standard virtualization software (e.g. VMware) can cope with libraries of such images and ensure that any mutation of the operating environment introduced by software installation can be undone.

Virtualization has the potential to simplify the preservation of reference workstations. One can imagine a pool of images shared among libraries to enable patrons of local public libraries full access to the FDLP materials. Virtualization does not solve one fundamental preservation issue – as the materials become obsolete, knowledge about how to use the required software becomes more obscure.

To partially address the issue of loss of application knowledge, we have experimented with techniques to automate mounting and installing an ISO image within a VMware virtual machine. The basic approach we are exploring is the creation of a single install application which responds to user “click” on ISO images by selecting a local VMware machine, mounting the selected ISO image, and running an image specific installer script. This process might be further expanded by executing specific helper applications. To determine the possible utility of this approach, we surveyed 100 ISO images containing Windows/DOS executables. Of

these, more than 50 required a multi-stage installation procedure prior to use.

Because of legacy use in software testing, virtualization solutions such as VMware provide scripting interfaces enabling automated control of a virtual machine. In our work we developed a Perl script which utilizes the VMware VIX API to start and stop virtual machines as well as mount and unmount CD-ROM images (VMware 2008). We use the “snapshot” capability of VMware to ensure that all modifications made by a user are erased and to guarantee that each user is presented with a virtual machine in a known state.

The guest OS is configured with key applications for document and media browsing (Microsoft Office 97 with compatibility updates, Adobe Acrobat, VLC, and a current release of Windows Media Player) to provide the user with a simple, easy-to-use environment for browsing legacy documents on the mounted ISO image. ISO images are mounted from the network using a standard AFS client to provide volume access to IU.EDU in the global namespace. In cases where an installation is required, a precompiled Windows executable unique to the image is copied from the AFS to the guest OS and run at startup to automate the install process. A standard “wizard” provides the user with the option to cancel the installation if desired. These executables are simple wrappers around macro-style Windows scripts, implemented with the cross-platform wxWidgets GUI library. They are easily maintained and can be trivially ported to additional client platforms.

Our proof-of-concept trial with 100 ISO images containing legacy installation executables in the top-level directory demonstrated an additional advantage of this approach. Of the 66 images requiring local installations, the majority were hard-coded to look for a physical device such as a D: drive where the CD-ROM would originally have been mounted. This limitation is readily accommodated in an emulated environment.

Image Access

As discussed previously, we preserve CD-ROM data in the form of “bit faithful” ISO9660 images which are supported by a well defined standard. In this section we consider three issues relating to ISO images – access to the files within the image, distribution of a shared collection of ISO images using OpenAFS, and the creation of bit faithful images.

Access Within ISO Images

ISO images are directly supported in many operating systems (BSD, Linux, OS X) and emulation tools (e.g. VMware). Thus, if the only goal is preservation, ISO images are a sufficient target. In our work, we are interested in making the CD-ROM collection more useful in virtual than physical form. This requires the ability of a server to access the contents of a large collection of ISO images.

Our first approach was to exploit the ability of Linux to mount ISO images on *loopback* devices. By suitable creation of file links and configuration of the automounter it is possible to make a collection of ISO images appear to be mounted as subtrees of the host file system. There are several limitations to this approach. Many of the files are in obsolete formats, many files contain links that implicitly depend upon the ISO image being mounted as a virtual CD, some of the ISO images were created from Macintosh computers and are not fully supported by Linux, and allowing web access to trigger kernel mounting events has scaling and possible security issues. Thus, we have moved to utilizing a widely available library, `libiso9660`, to enable direct access to the contents of ISO images without mounting.

As discussed above, some CD-ROMs created for Macintosh computers have compatibility issues. This issue is manifested in pairs of identically named files representing “resource” and “data” forks (a Macintosh concept). While the end user is generally interested in the data fork, Linux is unable to extract the correct file from an ISO images – indeed we had to patch `libiso9660` to “do the right thing.”

Finally, there are significant issues arising from standard ISO9660 extensions such as Joliet and Rock Ridge which were intended to overcome file naming and metadata issues in the original ISO9660 specification. These issues make it difficult to correctly render all file names and utilized these rendered names to find files in ISO images.

Image Distribution

ISO images are quite large – as much as 8 GB for recent DVD based titles – and in most cases only a small fraction of the information in an image is required either to determine that a title is of no further interest or to satisfy a specific data query. Thus, it is extremely inefficient to download entire CD-ROM images on demand. Utilization of most CD-ROM titles is extremely low and even with rapidly decline storage costs it doesn’t appear to make sense to mirror a large CD-ROM image collection at all libraries and furthermore, such widespread mirroring complicates the access control required to satisfy copyright restrictions. In this section we discuss our use of existing distributed file system technology, the Andrew File System (AFS) to enable sharing of an ISO image collection, with proper access controls, and in a manner that largely eliminates the need to copy ISO images to satisfy patron requests (OpenAFS 2008). Our use of AFS enables access to CD-ROM images through web-server based browsing, through remote mounting on workstations or emulators, and copying of entire images in the rare cases that might be necessary.

Key characteristics of AFS that we exploit are a global namespace, transparent storage migration, storage mirroring, multi-domain authentication using the widely deployed Kerberos protocol, flexible access control based on ACLs, and clients for most common operating systems. In the Emulation section we discussed a simple

application that automates mounting ISO images in VMware Workstation. Furthermore, our web-application accesses both its metadata and ISO images through AFS without apparent performance issues. An exception may be high bandwidth movies where performance is significantly improved by image copying. This level of performance stands in direct contrast to the issues observed with mounting ISO images as filesystems on a local server, including overhead on the kernel and scalability limitations.

In our prototype system, the ISO images are distributed across 5 volumes [file:///afs/iu.edu/public/sudoc/volumes/\[01-05\]](file:///afs/iu.edu/public/sudoc/volumes/[01-05]) which are accessible from anywhere by anybody. We separately maintain metadata in METS form <file:///afs/iu.edu/public/sudoc/metsxml> generated from the Indiana University Libraries MARC records and which provide links to the raw images. Our web-server uses Swish-E to index these METS records, and utilizes XSLT to format the METS records. We anticipate that in a production system libraries may integrate such a collection into their own catalogs or digital repositories by mining the METS records. While all of the FDLR materials are openly available, we have created an additional collection of materials (e.g. Unesco) which are subject to copyright restrictions and accessible only in the IU Kerberos domain.

The model we anticipate is one where a collective of libraries share responsibility for creation of metadata and ISO images and share these materials through a dedicated AFS domain. Individual libraries could contribute materials through a local volume server and could control access to these materials through access control lists (ACLs). Access by patrons of other institutions would be enabled by linking to these participating institutions' Kerberos domains and by appropriately managing ACLs. A key issue for such a collective will be the development of effective administrative policies and tools; ACLs provide an effect enforcement mechanism, but aren't sufficient. For example, suppose copyright restrictions required that a particular item be accessible by only one patron at a time. While this restriction could be implemented by modifying the appropriate ACL at access time, we have not created administrative tools to perform such modifications.

Image Creation

Bit-level preservation of CD-ROMs would appear to be relatively straightforward – organization of information is governed by a well defined standard (ISO9660 (ECMA 1987), the data are protected by error correcting bits which make it possible to detect and correct most errors, and many software packages exist for ripping ISO images which are standard files containing the raw data from the CD-ROM. However, our experiences in preserving more than 4500 CD-ROM images have uncovered several significant pitfalls at both the bit-preservation and application levels. These pitfalls fall into two major categories – poor conformance to the

underlying standards, and inappropriate contextual dependencies embedded in the preserved data.

Operating systems such as Linux, Unix, and Windows all treat CD-ROM drives as “block devices” in which the raw data can be accessed as a single large binary file organized in fixed-sized blocks. For CD-ROMs, these blocks are called sectors and typically consist of 2048 bytes of data with additional error correcting bits used by the drive hardware to detect and correct bit errors.

The operating system interprets the contents of this binary file to provide a *file system* view consisting of a tree shaped hierarchy of directories and files which can be accessed by applications through standard file operations such as *open*, *read*, *write*. The binary file is organized according to the ISO9660 standard (ECMA 1987) described shortly.

In principle, preserving the contents of a CD-ROM consists of copying this binary file (called an ISO image) onto another media. For example, Microsoft provide instructions for doing just this (<http://support.microsoft.com/kb/138434>) and most available Windows tools for creating ISO images appear to follow this basic procedure.

There are two significant problems with simply copying the bits off a CD-ROM. There is no obvious way to know that you have all the bits, and there is no way to know the bits that you have are all correct. The latter problem is ameliorated by the error correction bits on the CD-ROM which are utilized by the CD-ROM drive to detect and correct errors; however, for an archival copy this may not be sufficient. The problem of knowing that you have all the bits is complicated by the fact that CD-ROMs are typically created with additional blocks of zeros to assist in the physical process of extracting the bits that are part of the file system. Errors in reading these extra blocks are irrelevant. As we shall show, knowing that you have all the bits requires interpreting the underlying file system organization.

An ISO file system consists of fixed sized *sectors* organized in one or more volumes. Each volume begins with a dedicated sector, called a volume descriptor. This volume descriptor includes fundamental information such as an identifier, volume size (in sectors), sector size, and pointers to directory information (within the volume). The directory information includes *path tables* – a largely obsolete mechanism for quickly finding files, and a *root directory*. As with most file systems, directories are implemented as binary data structures embedded in ordinary files.

One approach to determining the amount of data within an ISO image is to find the volumes and compute the length of the volume from the volume header. This is greatly simplified by the fact that most published CD-ROMs (all in our data set) consist of a single volume. Unfortunately, the volume header information is frequently wrong – 19% of the CD-ROM images we created had incorrect size information in their headers. The number can both be too high and too low. One fairly benign case arises in “track at once” recording when

images are one sector shorter than advertised. The advertised length can also be too high when the recording software computes the image size prior to “compacting” the file system. This problem emerged in our work when we began experiments with file migration and found significant numbers of truncated files within ISO images we had created.

To circumvent the “bad header” problem, we wrote programs that walk the file system in an ISO image computing the starting sector and length of each file (including directories). Using this technique we were able to determine the “true” end of the image.¹ Unfortunately, approximately 10% of the images we created with Windows based software were truncated before the end of the image. Experiments with a variety of windows based tools on multiple machines confirmed this behavior. In contrast, the Linux tool `dd` enables copying all of the raw data from a CD-ROM. In an experiment with 86 CD-ROMs whose images were truncated by Windows, we were able to read the entire ISO image for 81 using `dd`. Of the remainder, 1 CD-ROM was physically cracked. Thus, we expect the rate of failure to read all the bits of CD-ROMs to be under 1% provided the right tools are used.

Once it has been established that an ISO image contains all the relevant bits, it remains to be determined if these bits are correct. Since the CD-ROM publications of the GPO provide no additional checksum information, the only viable approach is to compare at least two images created from different copies of a CD-ROM title for consistency. As discussed above, it is crucial that only the relevant bits in an image be compared (or checksummed) as there is significant potential for spurious errors. In the case of the GPO publications, comparing copies of CD-ROMS is further complicated by the use of different identification schemes in the various FDLP libraries (the SUDOC number system is not universally applied and offers significant opportunities for ambiguity). Thus, simply identifying two copies of the same publication may require significant effort. One strategy we have explored is generating checksums for the first 1 Mbyte of each image in our collection as a convenient hash value for determining whether two CD-ROMs are likely to be the same publication.

Discussion

The work presented here addresses fundamental access problems faced by institutions with legacy CD-ROM holdings. Our project complements and operates alongside existing frameworks without significant additional overhead. It uses interoperable metadata and low-cost open source tools, and further enables secure, flexible sharing of archival materials. These factors,

¹This works for interchange level 1 and 2 CD-ROMs because they require all files to be contiguous. Interchange level 3 appears to be rare since it is incompatible with most operating systems.

along with viable strategies for file format identification, flexible migration profiles, and emulation support for legacy environments, provide a blueprint for future success in handling these types of collections.

Systems such as Fedora (Petitot et. al. 2004), Greenstone (Greenstone 2008), and DSpace (DSpace 2008) provide an established basis for archival management systems. In our view, there are fundamental access and preservation issues with CD-ROM collections that these systems do not adequately address. Foremost is that CD-ROM collections typically consist of a comparatively small number of large objects (physical CD-ROMS and DVDs, or their bit-identical ISO-9660 images) that will generally see only fractional access. While the images as a whole contain a large number of disparate file types, these files are bound within the context of individual ISO images. Our approach emphasizes the inherent interrelationship of items within an ISO image over those between images.

This project explicitly enables libraries to build shared virtual collections using predominantly off-the-shelf, open source tools. The methods discussed in this paper address a number of outstanding access issues faced by institutions holding legacy digital materials, and may be readily integrated into existing infrastructure.

References

- Arms, C. and Fleischhauer, C. 2005. *Digital Formats: Factors for Sustainability, Functionality, and Quality*. Washington D.C.: IS&T Archiving Conference
- Brown, A. 2006. *Digital Preservation Technical Paper 1: Automatic Format Identification Using PRONOM and DROID*. UK: National Archives
- CD/DVD Database. 2007. *Regional Depository CD/DVD Database*. http://www.uky.edu/Libraries/CDROM_2004-_inventory.xls. Accessed August, 2007.
- Council on Library and Information Resources. 2000. *Authenticity in a Digital Environment*. Washington D.C.: Council on Library and Information Resources,
- DSpaceWiki. 2006. *JHOVE Integration with DSpace*. <http://wiki.dspace.org/index.php/JhoveIntegration>. Accessed November, 2006.
- ECMA. 1987. *Standard ECMA-119 Volume and File Structure of CDROM for Information Exchange*. <http://www.ecma-international.org/publications-/standards/Ecma-119.htm>. Accessed July, 2008.
- ERP5. 2008. *How To Use OOOD*. <http://www.erp5.org/HowToUseOood>. Accessed July, 2008.
- FDLP. 2006. *About the FDLP*. http://www.access.gpo.gov/su_docs/fdlp/about.html. Accessed November, 2006.

- FDLP Desktop. 2005. *2005 Minimum Technical Requirements for Public Access Workstations in Federal Depository Libraries*. <http://www.fdlp.gov/computers/mtr.html>. Accessed July, 2008.
- FDP. 2006. *CIC Floppy Disk Project*. <http://www.indiana.edu/~libgdp/mforms/floppy/-floppy.html>
- GDFR. 2008. *Global Digital Format Registry*. <http://hul.harvard.edu/gdfr/>. Accessed November, 2008.
- GPO. 2006. *Depository Library Public Service Guidelines For Government Information in Electronic Formats*. http://www.access.gpo.gov/su_docs/fdlp/mgt/pseguide.html. Accessed July, 2008.
- Heminger, A. R. and Robertson, S. 2000. *The digital Rosetta Stone: A Model for Maintaining Long-Term Access to Static Digital Documents*. Communications of AIS **3**(1es): 2. Atlanta, GA: AIS
- Hernandez, J. and Byrnes, T. 2004. *CD-ROM Analysis Project*. http://www.princeton.edu/~jhernand/Depository_CD-ROM_Legacy.ppt. Accessed December, 2006.
- Hoeven, J. R. V. D., Diessen, R. J. V., et al. 2005. *Development of a Universal Virtual Computer (UVC) for long-term preservation of digital objects*. Journal of Information Science **31**(3): 196-208. Delft, Netherlands: JIS
- JHOVE. 2006. *JHOVE - JSTOR/Harvard Object Validation Environment*. <http://hul.harvard.edu/jhove/>. Accessed July, 2008.
- Lawrence, G. W., Kehoe, W. R., et al. 2000. *Risk Management of Digital Information: A File Format Investigation*. Washington D.C.: Council on Library and Information Resources.
- Leonard, T. 2008. *The Shared MIME-info Database Standard*. <http://standards.freedesktop.org/shared-mime-info-spec/>. Accessed July, 2008
- Lorie, R. A. 2002. *A methodology and system for preserving digital data*. Proceedings of the second ACM/IEEE-CS joint conference on Digital Libraries 312-319.
- Maniatis, P., Roussopoulos, M., et al. 2005. *The LOCKSS peer-to-peer digital preservation system*. Transactions on Computing Systems **23**(1):2-50.
- McCray, A. T. and Gallagher, M. E. 2001. *Principles for digital library development*. Commun. ACM **44**(5): 48-54.
- Mellor, P., Wheatley, P., et al. 2002. *Migration on Request, a Practical Technique for Preservation*. ECDL '02: Proceedings of the 6th European Conference on Research and Advanced Technology for Digital Libraries: 516-526.
- Moore, R. 2006. "Building Preservation Environments with Data Grid Technology." <http://www.archives.gov/era/pdf/2006-saa-moore.pdf>
- MTR. 2005. *2005 Minimum Technical Requirements for Public Access Workstations in Federal Depository Libraries*. http://www.access.gpo.gov/su_docs/fdlp-computers/mtr.html. Accessed July, 2008.
- NatSciBoard. 2006. *NSF's Cyberinfrastructure Vision for 21st Century Discovery. Version 5.0*. http://www.nsf.gov/od/oci/ci_v5.pdf. Accessed November, 2006.
- OAI-PMH. 2006. *The Open Archives Initiative Protocol for Metadata Harvesting*. <http://www.openarchives.org/OAI/openarchivesprotocol.html>. Accessed December, 2006.
- OpenAFS. 2008. *OpenAFS*. <http://www.openafs.org/main.html>. Accessed May, 2008.
- Petinot, Y., Giles, C. L., et al. 2004. *A Service-Oriented Architecture for Digital Libraries*. International Conference on Service-Oriented Computing. New York, NY: ACM
- PREMIS. 2006. *PREMIS (PREservation Metadata: Implementation Strategies) Working Group*. <http://www.oclc.org/research/projects/pmwg/>. Accessed November, 2006
- PRONOM. 2008. *The technical registry PRONOM*. <http://www.nationalarchives.gov.uk/pronom>. Accessed July, 2008.
- Rabinowitz, J. 2004. *Simple Web Indexing for Humans – Enhanced*. <http://www.swish-e.org>. Accessed July, 2008
- Rothenberg, J. 2000. *An Experiment in Using Emulation to Preserve Digital Publications*. Netherlands: Koninklijke Bibliotheek.
- VMware. 2008. *VMware Server*. <http://www.vmware.com/products/server/>. Accessed July, 2008.
- Witten, I. H., Bainbridge, D., et al. 2001. *Power to the people: end-user building of digital library collections*. Proceedings of the First ACM/IEEE-CS Joint Conference on Digital Libraries: 94-103.
- Woods, K. and Brown, G. 2008. *Migration Performance for Legacy Data Access*. To appear in: International Journal of Digital Curation **3**(2).

Preservation Of Web Resources: The JISC-PoWR Project

Brian Kelly^{*}, Kevin Ashley⁺, Marieke Guy^{*}, Ed Pinsent⁺, Richard Davis⁺ and Jordan Hatcher^{\$}

^{*} UKOLN, University of Bath, Bath, BA2 7AY

⁺ ULCC, 20 Guilford Street, University of London, WC1N 1DZ

^{\$} Consultant, Opencontentlawyer.com

B.Kelly@ukoln.ac.uk, M.Guy@ukoln.ac.uk, E.Pinsent@ulcc.ac.uk, R.Davis@ulcc.ac.uk,
K.Ashley@ulcc.ac.uk, jordan@opencontentlawyer.com

Abstract

This paper describes the work of the JISC-funded PoWR (Preservation Of Web Resources) project which is developing a handbook on best practices and advice aimed at UK higher and further educational institutions for the preservation of Web sites and Web resources. The paper summarises the challenges institutions face in preserving Web resources, describes the workshops organized by the project in order to identify the challenges and identify appropriate best practices, and outlines areas in which further work is required.

Background

The preservation of Web resources is a topic that is of interest to many involved in digital curation issues. It presents many interesting technical challenges in terms of capture and access, and organisational and resource-oriented problems, some of which are shared with other aspects of digital preservation and some of which are unique to Web resources. How does one select material? When are we trying to preserve information and when is it the experience, behaviour or appearance that is paramount? How straightforward is it to move Web resources between curatorial environments? Most everyone knows that information persistence on the Web is a fragile thing. And, as Rusbridge has observed [1] even those who care about information persistence don't necessarily do a good job of it on their Web sites. This, despite the fact that good advice about URI persistence has been available for some time [2]. URI persistence is just one small (albeit important) part of the problem that illustrates the wider issues that surround Web preservation in an institutional context.

Not everything on the Web needs to be kept. And there's more than one way to go about keeping it - often it's just the information that needs to survive, and the particular way it is presented on a Web site today is not, of itself, worthy of long-term preservation. Yet there's a lack of knowledge where it's needed about **how** to preserve Web resources, and even when people know how to do it, for some reason it just doesn't happen. That's not a situation the scholarly community is comfortable with, which led to JISC funding the work which is the subject of this paper.

We describe a project funded by the JISC with the aim of producing a series of guidelines on the preservation of Web resources in UK academic institutions. The project,

JISC PoWR (Preservation of Web Resources), which is funded from April – September 2008, has established a blog [3] and is running a series of workshops which are helping to gain a better understanding of the challenges institutions face in preserving Web content and support the development of guidelines on best practices.

The paper summarises the work of the project to date, including two workshops which helped to identify challenges and strategies for addressing the preservation of Web resources in a managed Web environment and use of externally-hosted Web 2.0 services.

The project is taking a broad view of what constitutes a Web resource, and hence the remit of the guidelines we will produce. But not everything that is Web-accessible will be covered; for instance, University finance systems will often have a Web interface but are not themselves intrinsically Web resources. But access logs, intranets and externally-hosted content are certainly amongst the types of resource we have been considering, along with the externally-hosted Web 2.0 services which are of growing interest within the sector.

The workshops have endeavoured to bring together institutional stakeholders who might not otherwise encounter each other, such as records managers and Web managers. We are also conscious that it is important to separate decisions about what policy says would be ideal from what is achievable using current resources and technology. We want to bridge the gap between some of the information available about web archiving [16],[17] and their application in a wider organisational context. Where a decision is taken to preserve material, we intend to help institutions make sensible choices between in-house solutions, explicitly-outsourced solutions and what might be described as passive outsourcing: the belief that someone else will do the job for us.

The Preservation Challenges

The Drivers

There are many drivers for undertaking Web site and Web resource preservation within a higher educational

institution: institutional policy, legal requirements, and research interests are just a few.

The University is an organisation with business continuity interests that need to be protected. It will have an interest in protecting, managing and preserving certain types of Web content to meet legal requirements and manage its information legislation compliance. The JISC have pointed out that increasingly "*websites may be a unique repository for evidence of institutional activity which is unrecorded elsewhere, and this is often unacknowledged*" [4]. For audit purposes, for example, reference to archived copies of institutional Web sites may be required for the checking of strategic, legal, financial, contractual or scholarly information. If unique records are indeed being created, stored and published on the web, then we'll need to establish their authenticity as records, or as trustworthy and reliable versions of pages.

The University has a responsibility to staff, students, and researchers. Certain services for examinations and assessments are increasingly delivered on the Web; there are static resources accessed through the Web, such as library and learning materials. Students and staff are themselves creators of Web resources, in the form of wikis and blogs; these may sometimes generate content of lasting value. The Web site can be seen as a publication tool, or a dissemination tool; it may be governed by an agreed publication programme. Students will be making career choices, and staff will be making business decisions, based on information they find on the Web - and more importantly, when and where they found it. Does the University have a record of its publication programme? Can it roll back the Web site to a particular point in time to verify what was published two or three years ago? And does it need to be able to roll back the site itself, or the information resource behind the web site ?

Research interests are reflected in the increasing number of Web resources that have potential longevity and re-use value, a category that may include scientific research outputs and e-learning objects. Time, money and energy will be wasted if these resources are not preserved, or at the very least protected or managed in some way. There is a heritage dimension and this reflects the University's social responsibility to the academic community; viewed collectively, Web resources will provide interesting insights into the development of Higher and Further Education digital initiatives over the course of the last fifteen years.

Legal Challenges

Preservation of Web resources places the preservationist in a similar position to a publisher as the task can require copying of a resource. This activity, and the others of the preservationist, can carry some legal risks – many of the same risks as the creator of the resources faces in the first place.

Legal issues that can arise when preserving Web resources include:

- Freedom of Information (FOI) legislation, which entitles the public to request recorded information from public authorities, including universities;
- Data Protection Act (DPA) rules governing the use of personal information;
- Intellectual Property Rights (IPRs), particularly copyright;
- Criminal and civil laws that relate to the content of the resource, such as defamation, obscenity, or incitement to racial hatred;
- Contractual obligations such as Terms of Service (ToS) for third party Web sites, particularly in the Web 2.0 space (such as Facebook or Slideshare, mentioned below).

Naturally this list does not exhaust all of the potential legal issues, and each preservation project will have different risks and legal obligations. When examining the potential legal issues on a particular project, it might be useful to break down the issues into the following:

1. **Preservation of a resource because of a legal requirement.** This could be, as mentioned above in a records management context in order to comply with FOI legislation. The "legal requirement" area could be further divided into hard requirements – laws that say something must be retained or preserved – and soft requirements – self-imposed rules to avoid exposure to some legal risk. One example for a soft requirement might be keeping a copy of a Web site's terms and conditions as they evolve in order to prove what terms governed at each exact time.
2. **Legal requirements not to preserve a resource, such as the 5th Data Protection principle:** "Personal data processed for any purpose or purposes shall not be kept for longer than is necessary for that purpose or those purposes" - see [5].
3. **Preservation of content for a non-legal reason but for which legal issues must be addressed.** This could include any number of reasons, such as for cultural heritage.

The notion of **risk management** rather than absolute risk avoidance does however act as an overall umbrella to these three areas. Clearly rules that firmly require information to be retained or not must be complied with. Concentrating on the possibility of legal liability too much for every area in-between does run another kind of risk – losing the resource.

Engaging With The Communities

The first JISC PoWR workshop took place on 27th June 2008 at the University of London. The day was intended to

provide an introduction to the concept of Web preservation and to provide participants with the opportunity to discuss the technological, institutional, legal and resource challenges this presents. The workshop format comprised of a number of presentations and discussion group sessions.

The launch workshop had two primary aims: Firstly to bring together a number of different communities to whom Web resource preservation is of potential importance. This was achieved with an attendance of over 30 people from a wide range of professional groupings, including the Web Management, Records Management and Archives communities. Secondly, to obtain input into the main project goal: the creation of a handbook that specifically addresses digital preservation issues of relevance to the UK HE/FE Web management community. During the day this feedback was provided on the form of suggested content for the handbook, possible delivery scenarios for the handbook and discussion looking beyond the handbook.

The initial presentations and first breakout session explored the challenges that Web resource preservation presents. Consideration was given to the complex nature of the Web: both through its size, transience and reliance on technologies, many of which are external hosted. It was established that Web resource preservation is also hindered by confusion over whose responsibility it is and how decisions on selection should be made. Delegates agreed that one clear requirement for the handbook was the establishment of an effective driver to motivate management buy-in.

The need for fusing of different communities was well demonstrated in the case study presentation given by Alison Wildish and Lizzie Richmond from the University of Bath. Alison (Head of Web Services) and Lizzie (University Archivist, Records Manager and FOI Co-ordinator) described how when asked to give a presentation on their approach to Web resource preservation they had initially felt apprehensive. Although Lizzie could see the value in theory she felt that in practice it was *“too huge a task”*, while Alison admitted that she wasn't really interested and had asked herself *“why is it something I should think about now?”* The task of preparing their presentation, in which they considered the necessary activity of preserving the University prospectus, gave them an understanding of the need for a collaborative approach to the preservation of Web resources.

After lunch a presentation was given on the relevant legal issues Web resource preservation broaches and the suggestion was made that delegates shouldn't panic. A risk assessment approach should be taken and the danger of not preserving should be given a higher priority than legal quandaries.

The second breakout session required delegates to consider possible scenarios related to Web resource preservation. For example one scenario required participants to provide examples of how their organisation's Web site has developed since it was launched. Although there was a lot of 'folk memory' and anecdotal evidence (also known as tacit knowledge) most participants felt they would be unable to reproduce screenshots showing changes to their institution's home page and were forced to rely on third party services, such as the Internet Archive, to provide snapshots of pages on the institutional Web site.

The concluding presentation offered some constructive approaches to protecting an institution's Web site in the short to medium term as part of a records management programme. It was suggested that delegates identify their resources, collaborate with others who have an interest in this area, choose the appropriate approach (or approaches) and accept that the preservation strategy may not, at this stage, include everything. The feedback obtained from attendees during the day will aid in the creation of a blueprint to be given in the project's handbook for the preservation of Web sites and Web resources.

A number of resources were developed for the workshop including three briefing papers on preservation tips, mothballing Web sites and Creative Commons licences. The main presentations were made available via the project blog, with links to audio recording of the talks also provided [6].

Preservation in a Web 1.0 Environment

The Web Managers' Perspective

Sometime in the mid-90s, institutions everywhere seemed to have set up a Web service. At first the service probably contained just a few pages of contact details and institutional overview, although in others cases, departments and individuals may have been able to create their own content sites in sub-sites on a main departmental or institutional service.

Responsibility for managing the Web site may have originated in the Computing Services department, with people skilled in technologies such as HTML, Javascript and CSS. For them term "archiving" would mean creating TAR and ZIP files and painstaking management of sets of daily, weekly, monthly backup tapes. To the 'WebFolk' it was considerably less likely to mean "keeping a copy of the previous version of the Web site that we can look at again sometime in the future". This is unfortunate as those early Web sites will have been relatively easy to archive and preserve. By comparison with today's Web resources (which may make use of customisable portals, database-driven services, embedded applications, etc.) collecting a few directories of HTML and JPEG files will have been a

trivial task for IT professionals capable of setting up and managing the complexities of Web server software.

Since those early days the Web has grown in sophistication and in complexity. Expectations of design, user interface, content and functionality have grown, for external marketing and publicity services, internal information management on an Intranet and, especially in a Web 2.0 environment, for richer Web-based applications such as, Virtual Learning Environments (VLEs). The Web has now become the platform and interface of choice for virtually every kind of information system.

As we have discovered through our JISC PoWR workshops, Web managers are likely to see their main responsibility as being to their users – keeping online systems useful, usable and up-to-date. That alone requires a lot of running just to stand still. In addition to changing technology and standards, and ever greater demands from creators and consumers of information and publications, there is also an ever-changing regulatory and legislative environment, which may require a complete overhaul of the design of the system.

Therefore it is easy to see why issues that have been identified as key to effective Web preservation – things like persistence, continuity, accessibility, and preservation management – may not be prioritized, or, indeed, even recognised, by members of institutional Web management teams.

Content Management Systems can help with day-to-day management of the Web content. Many even provide version control, though it may be questionable whether such systems could easily recreate a reliable and authentic copy of not only a Web page, but also its environment, functionality, context and embedded external resources. Even if they did, does that commit us to using the same CMS, possibly even the same version, for as long as we want that feature?

What about other systems? Most Web Managers are probably happy to leave responsibilities for management of the content of Web-based institutional repositories, VLEs, discussion boards, etc, to those who requested them. Once again, there are backups, and if any content needs any special attention, each discrete system has a manager whose responsibility that ought to be.

There is just such a huge range of resources on the web it's more than enough to keep a typical Web manager and Web team busy, without them having to consider the nature of records and publications, preservation and archiving as well.

But, as James Currall pointed out [7] that it is simply not a legitimate problem to drop on "the Web guys" lap, any more than it is one that has an instant technological

solution. Deciding what to preserve, and why, is an issue of institutional policy, that needs to be addressed at a senior level across all departments and functions with a Web presence. In universities today, that means everyone.

Armed with a clear brief from policy, Web managers and developers can start thinking about how to capture selected Web objects, and work with the records managers to decide how to store, manage and make them accessible – and what the resource implications of these actions will be

Information Management

The JISC PoWR project proposes that approaches adapted from the information management professions - lifecycle management, records management, archive management - will help with some of the issues raised at the first workshop and discussed in the previous section. We must *manage* resources in order to preserve them (and equally, we must manage them in order to make auditable decisions *not* to preserve them.). An unmanaged resource is difficult, if not impossible, to preserve. Information lifecycle management, if adapted, can help manage Web resources. A records management approach will help to define preservation periods for business records or for legal reasons, even if permanent preservation is not required. Permanent preservation - usually the concern of an archivist - is usually only appropriate for a small subset of resources, for research or cultural purposes.

A records management approach, for example, may be considered suitable when it is known that a Web site contains unique digital records. The Web site itself could be viewed as a record, or - more likely - a potential place where records can be stored or generated. A records manager might ask if people (external and internal) are making business decisions, or decisions about their academic career, based on the information they find on the Web site. Or if transactions, financial or otherwise, are taking place over the Web site and whether the University needs to keep records of these transactions. Are there unique, time-based, evidential records being created this way? If so, how can we capture them?

A Web manager could co-operate with the records manager (and vice versa) to the extent that the site, or parts of it, can start to be included in the University Records Management programme. This may entail a certain amount of *interpretation* as well as co-operation. University policies and procedures, and published records retentions schedules, will exist; but it is unlikely that they will explicitly refer to Web sites or Web-based resources by name. Where, for example, institutional policies affecting students and student-record keeping are established, we need to find ways of ensuring that they extend their coverage to all appropriate Web resources.

The attraction of bringing a Web site in line with an established retention and disposal programme is that it will work to defined business rules and retention schedules to

enable the efficient destruction of materials, and also enable the protection and maintenance of records that need to be kept for business reasons. The additional strength is that the Web site is then managed within a legal and regulatory framework, in line with FOI, DPA, IPR and other information-compliance requirements; and of course the business requirements of the University itself.

The Challenges of Web 2.0

The second JISC PoWR workshop took place on 23rd June 2008 at the University of Aberdeen. This workshop was held as part of UKOLN's annual Institutional Web Management Workshop. The workshop took place after a plenary talk at the event on "*The Tangled Web is but a Fleeting Dream ...but then again...*" given by James Currall [7]. The talk helped to raise the profile of Web preservation for the 180 delegates at the event.

This workshop [8] lasted for 90 minutes. In this short time the discussions and recommendations from the first workshop were described. Participants were then given the opportunity to give their views on a series of scenarios based on use of Web 2.0 technologies including:

- Use of wikis
- Student blogs
- Repository services, such as Slideshare
- Use of Twitter
- Use of Skype
- "Amplified conferences"

The discussions on these particular technologies helped to inform the plans for guidelines on how to address the preservation challenges when making use of Web 2.0 technologies.

Some of the issues that were discussed with regard to these Web 2.0 technologies included:

Wikis: Examples were given of use of externally-hosted wiki services to provide user input, note-taking and user feedback at events. A number of wiki services had been used at a variety of events organized by UKOLN. Typically the wikis were open to anyone for creating and editing the content. This open access policy was taken in order to minimize authentication problems. The approaches taken to the longer term management of the content was to tighten up the access shortly after the event so that only registered users could edit the content. At a later date only the event organizers could modify the content. In addition the content was migrated from the third party wiki service to a managed environment on the UKOLN Web site.

Blogs: An example of an institutional student blogging service was discussed. Although use of an in-house system might be regarded as allowing the content to be

safely managed without the risks associated with use of third party services, there was discussion regarding institutional policies on the management of student data and accounts once the student has left the institution. An example was provided of a student blog which had been migrated from an institutional blogging service to a third party service once the student had left the institution [9]. This example illustrated some of the difficulties in migrating blog content, including bugs in export tools, the limitations of such tools (e.g. only exporting text, and leaving links to embedded content), the loss of blog comments or the difficulties in linking comments with the original blog posts and the difficulties of redirecting the address of the content to new services.

Slideshare: Slideshare is an example of a third party service used for sharing resources – in this case slideshows created by software such as PowerPoint. Although hosting slides on Slideshare has been shown to enhance access to resources [10] there may be concerns over continued access if the Slideshare service is not sustainable over a long period. One approach which has been taken has been to provide a master copy of the slides in a managed environment on the institution's Web site, and to ensure that the title slides and the metadata on the copy on Slideshare provides links to the managed resource.

Twitter: Although many felt that micro-blogging tools such as Twitter should be regarded as personal chat tools with no need for institutional preservation policies for their content, it was pointed out that several institutions have already established official Twitter communications channels [11]. In addition UKOLN made use of an official Twitter account to support its IWMW 2008 event, with this technology being evaluated as a possible tool in case of emergencies [12]. There may be a need to take a more managed approach to such technologies used in this fashion. Possible approaches to such management might include the generation of Twitter posts from a centrally-managed service or the harvesting of the RSS feeds from the Twitter service itself. However of more importance than the technical approaches will be to have an understanding of the purpose of the service and the development of preservation policies which reflect those purpose.

Skype: The term 'Web 2.0' is now being used to cover a range of technologies including many communications tools. Internet telephony applications such as Skype are now being regarded as Web 2.0 applications, especially when, as is the case with Skype, there are additional applications which integrate with Web services. Is there, then, a need to include such applications when considering how to address preservation of Web resources in a Web 2.0 context? A simple response would be to argue that not only is recording of Skype

conversations out-of-scope, the recording of telephone calls without permission may be illegal. However there is a need to consider use of messaging channels which are often provided by such applications. In addition from an institutional perspective it may be desirable to develop preservation policies for digital resources which cover a diversity of technologies and aren't restricted to Web resources as conventionally understood.

'Amplified conferences': Lorcan Dempsey coined the term 'amplified conference' to describe events "*are amplifying their effect through a variety of network tools and collateral communications*" [13]. The IWMW 2008 event provided an example of an amplified event, with the provision of a Ning social networking environment, use of Twitter (described previously), a conference back-channel, streaming video of the plenary talks and videos of various informal activities surrounding the event. The variety of technologies which can be used to enhance the effectiveness of an event and increase its impact will provide particular challenges for the preservation of the associated resources. The approaches taken at the IWMW 2008 event have been to (a) document the third party services used, which also supports the event's approach to risk assessment [13]; (b) migration of appropriate data to managed environments; (c) provision of a diversity of services; (d) use of recommended tags to allow distributed data to be aggregated; (e) recording use of software in cases in which the long term sustainability may be questionable and (f) encouraging use of Creative Commons licence at the event to minimise legal barriers to reuse of the content.

Best Practices for A Web 2.0 Environment

We have described some of the approaches which are being taken to try and address the preservation challenges for an event which is seeking to be innovative in its use of Web 2.0 technologies. But it is acknowledged that the approaches which are being taken by early adopters will not necessarily be easily adopted for use by others. There is a need to document the underlying principles and illustrate how these principles can be implemented.

Why Preserve in a Web 2.0 Environment?

The two main questions which need to be addressed in a Web 2.0 context are the same questions which are relevant in a Web 2.0 environment: "*Why preserve?*" and "*What are you seeking to preserve?*". However the diverse ways in which Web 2.0 technologies are being used means that such questions may be more challenging. As we have seen the use of personal and social technologies to support institutional business processes is adding additional complexities to the preservation challenges. And with the diversity of services which are now available and being used for which we cannot guarantee long term sustainability there is a need to be clear as to whether we

are seeking to preserve the underlying data, the services used by the institution to fulfill its business processes or the end user experience. There is also the question as to whether it would be acceptable for Web 2.0 services to be lost – a question which may not be understood in, say, a financial context, but may be relevant if services are being evaluated in teaching and learning or research contexts. After all we cannot guarantee that Google will continue to provide a search service, but there are industries which have built services assuming that this will be the case.

Approaches to Preservation in a Web 2.0 Environment

Once the fundamental questions of "why?" and "what?" have been addressed there will be a need to answer the question of 'how?'. However rather than addressing the specifics of how for particular services some general principles are given below:

Data export: Can the data be exported from the service? Can the rich structure be exported? Can the data be exported in formats which can be imported into other applications or services?

Data import: Can the data be imported into new applications or services? Has the data export / import process been tested? Is any data lost? Do imperfections in the data cause migration difficulties?

Quantifying the costs of migration: What are the predicted costs of migration of the data? How will the costs grow if large-scale data migration is needed?

Content syndication: Can the content be syndicated (using technologies such as RSS or Atom) to allow the content to be made available in other environments?

Sustainability of service: Is the service likely to be sustainable? Are changes to the service likely to be managed gracefully?

Acceptance of risks of loss: Would you organisation be willing to accept the risks of loss of data or a service?

Risks of not using a service: Would you organisation be willing to accept the risks of not using a service (i.e. the missed opportunity costs or the costs of developing or purchasing an alternative service)?

Providing a diversity of content: Is it possible to provide a diversity of content, to spread the risks of data loss?

Embedding the learning: The key purpose of a Web 2.0 service may not be the data or the application itself but understanding the underlying processes. The purpose of the service may be complete after the learning has been embedded.

Risk assessment /management: There is a need to develop and share best practices and approaches to risk assessment and risk management.

Raising awareness: There is a need to raise awareness of the importance of preservation strategies.

What Next?

In many respects the challenges of preservation in a Web 2.0 environment have many similarities with preservation in a managed Web 1.0 environment: in both cases there are requirements to clarify why preservation is needed and what aspects of a service need to be preserved. Content managed within the organisation using a Content Management System may appear to be more stable, but we know that Web pages and, indeed, Web site domains, do disappear even from managed institutional environments.

The uncertainties in relying on use of third party services, especially if there are no formal contractual agreements, would appear to make use of Web 2.0 services a risky proposition. But on the other hand since many Web 2.0 services make it easy for content to be created and reused we may find that Web 2.0 services provide a better environment for preserving Web content.

This tension between technologies and approaches which meet immediate business needs and those which best meet long-term policies on information management and retention, is not specific to the web. But the speed with which web services are emerging and evolving make effective decision making more difficult and more urgent than has been the case with other IT developments. Helping institutions define clear, technology-neutral policies and then helping them apply those policies rapidly to emergent systems will be a key success criteria for the guidelines we are developing.

We are also aware that the guidelines may identify a niche for external service provision for the preservation of some web resources. Institutions cannot do everything for themselves; projects such as UKWAC [15], whilst demonstrating the economies of scale that can be achieved in Web archiving, preserve only what their curators select. A number of external service providers exist for web archiving [18] [19] but use of these services by PoWR's target community is vanishingly small. There are a number of possible reasons for this - lack of awareness, cost and an inappropriate service model being amongst them - yet the project has already identified a desire for services broadly like this. Understanding the scale of this requirement for third-party preservation and the ideal service provision model is outside the scope of what JISC PoWR can achieve today.

There is a need for Web site technologies and management tools to provide better ways of providing long term access

to resources, which will include decoupling the address of resources (URIs) from the technologies used to deliver those resources.

But perhaps of even greater importance than technological developments is the need for improved dialogue and shared understanding amongst those involved in developing and implementing policies on Web site preservation.

Life After JISC PoWR

The JISC PoWR project will deliver a handbook on advice and best practices for Web site preservation in an institutional context. But what is the future for Web site preservation after the project's funding ceases? Feedback from the workshops has already encouraged us to view the handbook as a living document, probably hosted on a wiki, rather than as a static publication. This will help to ensure that content remains relevant, although it is no guarantee of continued maintenance.

JISC already ensures that its funded projects are required to document their approaches to the preservation of project resources after the project funded ceases. Recommendations have been made previously by the JISC-funded QA Focus project, and a simple 'Mothballing Web sites toolkit' was developed [20] to help projects in identifying the policy and technical decisions they would need to make. It might be timely to revisit the development of a more sophisticated toolkit which recognised that projects are likely to make use of Web 2.0 services and ensured that projects had considered the preservation aspects of use of such services.

For institutions, it will be interesting to see whether different approaches to web resource preservation are equally effective and easy to implement. The project will not last long enough to examine this in depth.

Although the work of the JISC PoWR project has focused on the preservation policies and strategies which institutions should be developing there is also a need to consider the external changes that might be necessary in order to help institutions meet their needs in the most effective manner. The dialogue that the project has enabled between its partners has been fruitful and enlightening for all of us, and it has been rewarding to see similar bridges being built across professional divides as a result of the workshops. We hope that the project's longer-lasting outputs will help to sustain these links and build upon them.

References

- 1 Rusbridge, C. (2008) *RLUK launched... but relaunch flawed?* 21 April 2008, Digital Curation blog. <<http://digitalcuration.blogspot.com/2008/04/rluk-launched-but-relaunch-is-flawed.html>>

- 2 W3C (1998) *Cool URIs Don't Change*, <<http://www.w3.org/Provider/Style/URI>>
- 3 JISC PoWR (2008) JISC PoWR blog, <<http://jiscpowr.jiscinvolve.org/>>
- 4 JISC (2008) *JISC IIT: The Preservation of Web Resources Workshops and Handbook*, <http://www.jisc.ac.uk/fundingopportunities/funding_calls/2008/01/preservationwebresources.aspx>
- 5 JISC Legal Information Service (2007) Data Protection Overview <<http://www.jisclegal.ac.uk/dataprotection/dataprotection.htm>>
- 6 JISC PoWR blog (2008) *Workshop 1 - Resources available*, 30 Jun 2008, <<http://jiscpowr.jiscinvolve.org/2008/06/30/workshop-1-resources-available/>>
- 7 Currall, J. (2008) *The Tangled Web is but a Fleeting Dream ...but then again...*, Presentation at IWMW 2008, <<http://www.ukoln.ac.uk/web-focus/events/workshops/webmaster-2008/talks/currall/>>
- 8 Guy, M and Kelly, B. (2008) *Approaches To Web Resource Preservation*, IWMW 2008, <<http://www.ukoln.ac.uk/web-focus/events/workshops/webmaster-2008/sessions/guy/>>
- 9 Kelly, B (2007) *A Backup Copy Of This Blog*, 19 July 2007, <<http://ukwebfocus.wordpress.com/2007/07/19/a-backup-copy-of-this-blog/>>
- 10 Kelly, B (2008) *How Plenary Speakers Are Maximising Their Impact*, 18 June 2008, <<http://ukwebfocus.wordpress.com/2008/06/18/how-plenary-speakers-are-maximising-their-impact/>>
- 11 Kelly, B (2008) *The Open University's Portfolio Of Web 2.0 Services*, 3 July 2008, <<http://ukwebfocus.wordpress.com/2008/07/03/open-university-portfolio-of-web-20-services/>>
- 12 Kelly, B (2008) *Use of Twitter to Support IWMW Events*, 30 July 2008, <<http://ukwebfocus.wordpress.com/2008/07/30/use-of-twitter-to-support-iwmw-events/>>
- 13 Dempsey, L (2007) *The amplified conference*, 25 July 2007, <<http://orweblog.oclc.org/archives/001404.html>>
- 14 UKOLN (2008) *Risk Assessment For The IWMW 2008 Web Site*, <<http://www.ukoln.ac.uk/web-focus/events/workshops/webmaster-2008/risk-assessment/>>
- 15 Bailey, S and Thompson, D (2006) *UKWAC: building the UK's first public web archive*, Dlib 12(1) <<http://www.dlib.org/dlib/january06/thompson/01thompson.html>>
- 16 Brown, A (2006) *Archiving Websites: A Practical Guide for Information Management Professionals* London: Facet
- 17 Masanes, J (Ed.) (2006) *Web Archiving* Berlin: Springer Verlag
- 18 Hamzo (2008) *Hanzo: The web archiving company* <<http://www.hanzoarchives.com/>>
- 19 Archive-it (2008) *Archive-It home page* <<http://www.archive-it.org/>>
- 20 UKOLN (2005) *Mothballing Web Sites toolkit* <<http://www.ukoln.ac.uk/qa-focus/toolkit/mothballing-01/>>

Preserving the content and the network: An innovative approach to web archiving

Amanda Spencer

The National Archives
Kew
Richmond TW9 4DU
Amanda.Spencer@nationalarchives.gov.uk

Abstract

Government's use of the Web has required new approaches to Web resource preservation. The National Archives' approach draws on its experience of Web Archiving, as well as expertise in the live Web arena. By harnessing these two elements, The National Archives hopes to deliver a truly innovative user-centric service predicated on preserving the content of websites as well as utilizing the value of the Web as a network.

Introduction

The proliferation of websites in the workplace has touched every sector, and Government has been no exception. Since the early 1990's the Government has been using websites to present information: official reports, papers, transcripts of speeches, guidance, announcements, press statements, regulations and advice. The benefits offered by these new technologies means that services are increasingly being delivered via electronic means and through digital channels.

The evolution of websites, coupled with the size and ever-changing nature of Government, mean that these sites are vulnerable to technological problems, such as documents 'falling off' sites, or links being broken between resources.

The prevalence of broken Web links impacts negatively on the reputation of government because it is perceived that government is managing its information poorly; a frustrating user experience on line also has the potential to reduce public confidence, and parliamentary scrutiny of

This article, Preserving the content, and the network: An innovative approach to web archiving, was written by Amanda Spencer of The National Archives. It is published with the permission of the Controller of HMSO and the Queen's Printer for Scotland.

government is impaired by its inability to refer to key government documents.

This state of affairs was brought into sharp relief by Cabinet Ministers looking for documents on government websites, only to find that these documents had been moved or removed. On 19 April 2007 the leader of the House of Commons wrote to the incumbent Cabinet Office Minister expressing concerns over the issue of documents and information disappearing from websites, concerns supported by a sample survey of URLs (Uniform Resource Locators) cited in Hansard in response to parliamentary questions¹. It was noted that such links in the parliamentary record often failed to resolve.

As a consequence the Archiving Digital Assets and Link Management working group was formed in May 2007. The working group was comprised of members drawn from The National Archives of the UK, the British Library, Information Services at the House of Commons (formerly the Parliamentary Library), the Parliamentary Archives, and the policy unit at Central Office of Information.

Research

The working group identified a number of interrelated issues, supported by a number of pieces of applied research, which were all contributory factors to the loss of significant official information over time.

The current situation

The working group identified that the government has effective strategies in place for ensuring that all information laid before Parliament is published appropriately and flows through to The British Library, for long-term access and preservation. In the UK the British

¹ Preliminary research conducted by the then House of Commons Library (now Information Services at the House of Commons)

Library is custodian of a body of official government publications/information which has been built up over past centuries, for historical access and long-term preservation of government activity. Researchers and historians expect such long-term access to official information through the preservation work of British Library. These government strategies rely on the existence of a printed rendition and a degree of centralisation and control of these official publishing arrangements under the auspices of Her Majesty's Stationery Office (HMSO), which operates from within The National Archives. Where material falls outside such practices, there are no agreed procedures for ensuring that online information is preserved over historical periods and made accessible. At present e-government information is received and preserved by the British Library on a voluntary basis. An understanding of how Government handles this body of information was required.

Broken links

The first issue, and the primary driver for the establishment of the working group, as described above, was that of a breakdown in online access to information through links, highlighted by the preliminary findings of the Information Services at the House of Commons and confirmed by the research on Hansard conducted by The National Archives. A longitudinal survey of URLs cited in response to Parliamentary Questions and recorded in Hansard revealed that 60% of links in Hansard, cited between 1997-2006, had since broken, resulting in '404 Page Not Found' errors, suggesting that many government departments do not consider the issue of long-term access to government information.¹ And yet ministers and other government officials assume that the information situated at any given URLs cited in response to a Parliamentary Question will remain available in perpetuity.

Government's use of the Web

This issue is compounded by the fact that much of this information is increasingly only available electronically [e.g.s], not in print, and even then is not always filed in electronic document and records management systems (EDRMS) making the integrity of Web links crucial to the business of government. Some government departments,

¹Unpublished research conducted by John Sheridan at the Office of Public Sector Information (part of The National Archives since 2005) on Hansard which revealed that 60% of Web links cited in Hansard 1997-2006 are now broken, suggesting that many government departments do not consider the issue of long-term access to government information.

tend to post documents and information on websites in HTML, rather than PDF or Word, making it more difficult to extract and archive the stand-alone documents from the websites. Additionally, as our understanding of the potential use of the web has developed, there are powerful arguments in favour of using HTML instead of document formats such as PDF on the web. In terms of data mashing and the semantic web, HTML can yield far greater benefit than PDF, which can lock-in information and prevent its reuse for other purposes. Further, some Web-based database-driven content is only available via a Website's search interface. As a consequence any solution to the problems identified needed to take account of the changing nature of and potential uses of the Web.

Website Rationalisation

A further area of concern related to The Transformational Government Website Rationalisation programme. The Website Rationalisation programme, aimed at streamlining the Government's estimated 2,500 Websites², began in 2007 and is due to complete in 2011, and is concerned with delivering a better web user experience for the citizen seeking to access government information. Much citizen-focussed content will be converged onto the Directgov supersite. Other content may move to Departmental corporate sites and many websites will close in order to reduce government website proliferation. Although the issue of broken links is not a problem initiated or caused by the Website Rationalisation programme, there is a concern that it may exacerbate an already poor situation. It was agreed that there needed to be a policy and process for archiving and also for link management to ensure that information remains findable.

Given this state of affairs the working group concluded that all web-based information should be treated as an important contribution to the body of government information, and in particular that all online information that has been cited should remain available and accessible in its original form. This idea reflects an acknowledgement that the web has changed user behaviour in the way information is accessed.

The Options

The group explored a number of different options including improvements to existing practices, which would

² Government on the Internet: Progress in delivering information and services online, 29 April 2008, <http://www.publications.parliament.uk/pa/cm200708/cmselect/cmpublic/143/143.pdf>

involve re-issuing guidance on legal deposit of e-only publications, and using Digital Object Identifiers (DOIs). Both these options were considered to have certain drawbacks: guidance alone was unlikely to have significant impact, and if it did improve voluntary deposit, it was not considered to be very user- nor web-centric.

Using DOIs was initially a very popular idea, as it has been successfully implemented in the scientific publishing world. However, it was acknowledged that this system requires a high degree of centralised management and control on an ongoing basis, which was recognised as being too complex to achieve within a UK central government context. In working through these options the working group arrived at a solution which encompassed both elements: further guidance on managing websites and web content, and the principles behind use of DOIs and the idea of identifiers. With the existing web archiving programme at The National Archives we knew that it was possible to capture website content, and with the Website Rationalisation programme underway, we had already made a commitment to capturing more content, as websites closed or content was moved. Each piece of information already has an identifier in its URL, and in the web archive the same piece of information has a predictable archive URL, based on the original reference. For example, the most recent available copy of <http://www.mydepartment.gov.uk/page1.html> becomes [http://webarchive.nationalarchives.gov.uk/*/http://www/mydepartment.gov.uk/page1.html](http://webarchive.nationalarchives.gov.uk/*/http://www.mydepartment.gov.uk/page1.html). We just needed a way of matching an original URL with the archive version in the web archive. Because the European Archive identifiers are predictable we concluded that using a redirection component could enable us to run a DOI-type scheme using the web archive.

Following informal consultation of stakeholders, The National Archives (now leading the Knowledge and Information Function across Government) assumed responsibility for delivering the solution devised by members of the working group. This paper will outline the findings and outputs from this significant piece of work.

The National Archives and the Web Continuity Solution

The scale of the programme, the issue of trying to preserve both the content and the network, and the need for content capture to be as comprehensive as possible, required a truly innovative approach.

The project necessitated new thinking in web archiving to address a number of different, difficult elements:

1. How to capture significant levels of important Government information from possibly thousands of distributed, heterogeneous websites (including websites closing as part of the Website Rationalisation Programme);

2. Methods to ensure not only a greater capture of content, but also increase exposure of this content to the web harvesting crawler, from sites that vary hugely in nature;
3. Ensuring that links persist to ensure that users will always find the last available version of the page, whether it is on a live site, or in the web archive.

The extensive scope of the project has required a mechanism for auditing the Government web estate, for identifying and controlling the number of Government websites in operation, and for seeding the harvesting process. As a consequence new processes and tools have been developed. A central SQL Server database has been built for use as a registry of all UK Central Government websites. Originally intended solely as a means of seeding the harvesting process, discussions with other government stakeholders identified a need for a single source of up-to-date information about the live government web-estate, details of all websites, current and inactive, any schedules for content closure or convergence as part of the Website Rationalisation programme, and evidence of compliance with government web standards guidelines (such as the accessibility standard). The database will be available to all website managers in central government and the responsibility rests with them to keep their information current. Appropriate access controls have been applied so that website managers can only edit their own departmental records. In respect of Website Rationalisation, only the Transformational Government team at Central Office of Information, with responsibility for new government domain registration and as the Data Quality Officer for reporting on progress of the programme, will have 'update' access to the scheduled website closure and convergence dates.

The archiving of government websites is to be carried out using the most popular method of capture for large-scale programmes, remote harvesting using a Heritrix web crawler. The National Archives web archiving is carried out under contract to the European Archive, and the Web Continuity Project has meant a significant increase in the number of websites captured, moving from a selective archiving programme to a comprehensive programme involving all websites of central government departments, agencies and Non-Departmental Public Bodies (NDPBs). The research conducted by the Digital Assets working group which highlighted that often websites are the only source for particular documents, has required that the archiving programme recognize that the partial archiving of websites, often a result of the limitations of current remote harvesting technology, is not an adequate solution. As a consequence The National Archives has explored the possibility of using the XML sitemap protocol¹, to ensure

¹ <http://www.sitemaps.org/protocol>

that capture of the Government web estate is comprehensive.

The widespread adoption of XML sitemaps by government departments will have other associated benefits, most notably relating to web resource discovery using search engines.

Citizens increasingly use search engines to look for information hosted on websites on a wide variety of subjects. Government information forms a part of the enormous mass of information available, but if it is not exposed somehow to search engines indexing, it can be 'buried' among the mass of other, less relevant information, or worse, remain completely undetected, and therefore, unable to reach its intended audience. Search engine providers build indexes of available (i.e. exposed or linked to) information available on the World Wide Web. They are unable to include unlinked to or 'hidden' content (the so-called 'hidden' or 'deep' Web). Hidden content not only includes databases which can only be interrogated by queries, but also content which is essentially generated 'dynamically' or 'on the fly.' XML Sitemaps enable website owners to expose hidden content if appropriate, and moreover, allow website owners to have better control over what parts of their website they expose to search engines.

Software used to 'crawl' websites remotely in order to take archival snapshots operates in a similar way to search engine software. This type of crawling is the most efficient, robust and therefore widely used in large-scale crawling programmes. However, fundamentally, it can only crawl (and capture for archiving) content which is linked to, or exposed in some way. XML Sitemaps also enable website owners to expose hidden content if appropriate, to web archiving crawlers

The pan-government search group has recognised that more action is needed to ensure that current government information on websites is findable for citizens. It has also been recently recognised that action is needed to ensure that continued access to information over longer periods of time is also required. The National Archives through the Web Continuity project is developing a solution to the latter, which involves more comprehensive archiving of websites within the central government domain, and a method of links persistence so ensure that instances of 'broken' links to government information (acutely represented by 'broken' links in Hansard) are reduced.

The close relationship between searching for live context and capturing greater archival content and the recognition that both situations can be greatly improved through the adoption across government organisations of XML sitemaps, has meant that The National Archives has assumed responsibility for the Sitemap Implementation Plan across government.

Some people operating in the government Website arena already understand and use sitemaps, some know little or nothing about them, while others may understand what they are, but have little knowledge or experience of how to set about using them. Given that current knowledge and understanding of sitemaps and their practical uses is varied across government, The National Archives approach to sitemaps implementation is three-fold and is detailed below.

Online Instruction Packages (Breezos)

These packages will raise awareness of sitemaps and are designed to reach a wide non-technical audience. They have been written by a Third Party provider and comprise three separate modules:

- Introduction – why sitemaps are important and why you should have one
- Detail – what a sitemaps is
- Practical – How you can create a sitemap

The practical module will be complemented by research, testing and guidance organised by The National Archives (outlined in the following section).

Software

The National Archives has contracted a third party to evaluate a survey of the sitemap generation software market, against a set of pre-defined minimum functional requirements. Software vendors have been approached to validate their software against a rigorous set of technical and assurance-related claims, using the government-endorsed CESG Claims Tested (CCT) Mark scheme. It was intended that this exercise would be repeated every two years to ensure that the market-place evaluations remain timely, that new releases of software would be validated appropriately, and that only software which had been successfully validated using the scheme would be recommended to government organisations. However, there have been certain limitations with this aspect of the project, the primary issue is one of supplier incentivisation. The Claims tested scheme relies on suppliers financing the claims testing process, which costs around £20,000 per product tested. Few third party software suppliers seem interested in investing such a significant sum when the returns are likely to be relatively small. These products seldom cost more than £30.

Guidance

Guidance will be made available, which provides all the information necessary for installing sitemaps generation software, creating a sitemap, and deciding what to include or exclude from a sitemap.

Practical Implementation

The National Archives is working with its web archiving partner, the European Archive to ensure that where sitemaps are deployed, maximum effectiveness in capture

of content is achieved. After some discussions it was considered inappropriate to use the organisation's sitemaps as the primary mechanism for seeding the crawl. European Archive were concerned that if the sitemaps were out of date, or incomplete then the quality of the archived website instance would be impaired. Instead, agreement was reached that the sitemaps would be used to complement and enhance the content captured via the initial gather by the Heritrix crawler.

The relatively large-scale nature of the programme also favoured an automated approach to both the seeding of the crawls and the capture of preservation copies of the archived websites, which The National Archives receives under contract from the European Archive. As a consequence two interconnected workflows have been designed: one for the harvesting lifecycle which drives the crawling process and ends with the production of publicly accessible copies of the archived websites, made available by European Archive at a TNA IP address, and another which begins with the preservation copies being made available for ingest into The National Archives Digital Object Store (DOS)¹.

Automated harvesting begins with series level creation for cataloguing purposes, which provides another means of identifying websites and instances of websites within our wider collection, and enables websites to be situated within the context of the other digital and paper records of the creating department. The cataloguing information is captured in the website database, and is passed through to the European Archive so that it can be provided, alongside other metadata at ingest. The step-by-step harvesting process is dependant on a series of messages exchanged via FTP between the European Archive and TNA. Once this process is complete, a further message triggers the start of the ingest workflow and allows for the ingestion of multiple preservation copies in a single process step. Apart from the obvious benefits of having preservation copies stored separately from the presentation copies and their back-up versions, ingest into the DOS allows for active preservation of websites alongside the active preservation and migration of other digital records at The National Archives.

The final element of the new web archiving process concerns the use by government organisations of a redirection software component. Installation of this component will ensure the persistence of Web links and creates a different purpose for the web archive. The components, to be supplied by The National Archives,

Seamless Flow Newsletter Issue 1,
http://www.nationalarchives.gov.uk/documents/newsletter_issue1.pdf

following configuration by The Stationery Office (TSO), utilise open-source software and have been designed to work with Microsoft Internet Information Server versions 5 and 6, and Apache versions 1.3 and 2.0, which are the platforms most commonly used in UK Central Government². The IIS component is produced by Ionics www.codeplex.com/IIRF and the Apache component is the mod-rewrite module: http://httpd.apache.org/docs/1.3/mod/mod_rewrite.html

The component works by redirecting the user to the UK Government Web Archive in the event that the page could not be found (a 404 error). It does not replace any existing redirections on the live website. The component is also installed on the web archive site. Here its role is to rewrite the URL for the original department website, if it is not found in the archive. In this case, a further 301 is sent back to the requester. The departmental server will be reconfigured to recognise this URL as indicating that the archive has been checked, and will therefore be able to issue the appropriate custom error page.

- The user requests a URL e.g. <http://www.mydepartment.gov.uk/page1.html>
- If the URL is resolved, it is served back to the user in the normal way
- If it is not resolved, the web archive is checked to see if the page exists there. This is achieved by parsing the live URL into the predictable URL pattern used by the European Archive, e.g. http://webarchive.nationalarchives.gov.uk/*/http://www.mydepartment.gov.uk/page1.html
- If it the page is found in the web archive, the user is served with the latest version held there
- If the page does not exist in the archive, the user is served a "custom 404" from the original department website, stating that the page was not found on the original site, or in the archive.³

² Research conducted in December 2007, surveying 1101 Central Government Websites identified by Central Office of Information (COI) as part of phase 1 of Website Rationalisation, revealed the following usage: 644 uses of Microsoft IIS (of which 257 were using IIS 5.0 and 455 were using IIS 6.0); 287 users of Apache (of which 92 were using 1.3 and 76 were using 2.0)

³ For a diagrammatic expression of this, see Appendix 1, created by Brian O'Reilly at The National Archives in April 2008.

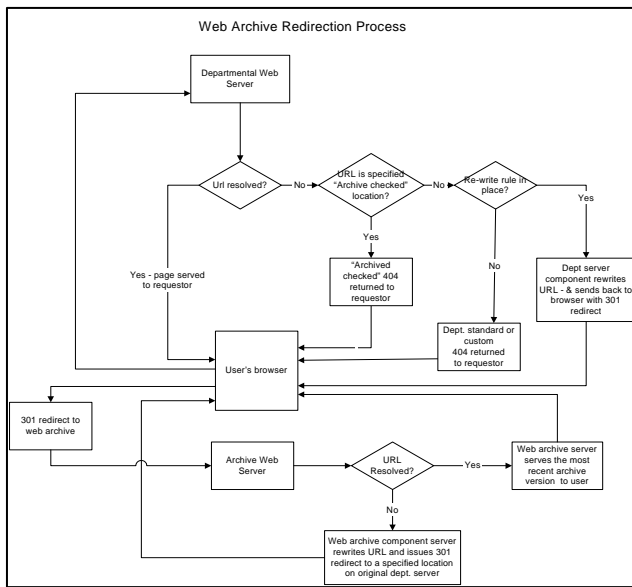


Figure 1. Diagram of Web archiving process,

The components configured, are, of course, only one means by which departments can choose to implement the required behaviour. However, they should be suitable for most government web server platforms. In order to provide guidance to those areas not using one of the two most popular environments, or where the configuration of their environment is atypical, documentation will be provided which describes the behaviour required as well as the technical means of achieving that behaviour. The technology developed, will however, only work successfully if government always uses Content Management Systems which allow the content to be published with persistent URLs. If URLs are randomly generated each time, or if Session IDs become part of the URL itself, then such URLs are irretrievable and will not match anything captured in the Web Archive.

The components are currently being tested at The National Archives, who run a load-balanced Microsoft IIS 6.0 web environment, and the Ministry of Justice, who use an Apache 1.3 environment. Load testing simulating up to 60 concurrent users has been applied, with favourable results. In order that the European Archive can also cope with increased demand for the web archive due to a much greater number of redirections, EA have introduced a mirrored infrastructure capable of failover, with a primary datacentre in Paris and a secondary datacentre in Amsterdam. The European Archive have also developed new indexing techniques to ensure that user requests for

the last archived instance of a website is locally cached, in anticipation of a greater number of calls for the latest snapshots predicated on the increased number of requests to the web archive arising from redirection.

The Web Continuity project and its use of the redirection component signals a marked departure from traditional web archiving programmes in the sense that it is not only concerned with preserving websites for their historical value, but also for their value as recently published information, and for their value in preserving the integrity of the network as a whole. The use of redirection software to persist links to the web archive implies the bringing together different audiences - the archival researcher and the user of current or semi-current information, and in doing so introduces the web archive to new communities of users, and introduces a temporal dimension to the web, which has implications not only for web archiving, but for the wider web more generally, ensuring a greater longevity of web pages than commonly experienced.

This has both benefits and drawbacks: the persistence of links, naturally has an immediate user benefit in that the user journey is less likely to be abruptly ended by a 404 message, but even more than this, the network of interlinking pages that makes up the World Wide Web is preserved, ensuring greater findability of content. However, the persistence of links to semi-current or even out-of-date information also has potential risks, and signals the need for information to be managed differently. Some government organisations have raised the issue of the potentially harmful effect that obsolete information could have if, for example, advice or guidance is revised and moved to a new location. Web users who still have the 'old' URL could potentially unwittingly access information that is no longer current, or which is actually completely inaccurate, for example, where new medical thinking has emerged. To mitigate the risk of archived information being mistaken for live information, The National Archives is working with the European Archive to develop a stripe located above the archived Web page. This stripe will be red in colour, will bear The National Archives logo, and will contain wording to the effect that the Web page is an archived snapshot, taken on a particular date. Various approaches to achieve this design have already been developed, the first of which used Frames. The use of Frames was considered inappropriate because of the issues it poses for accessibility. An iFrames solution was developed by Web developers at TNA, but testing with the European Archive revealed that the iFrame sat awkwardly with the layout of some of webpages already captured in the collection:

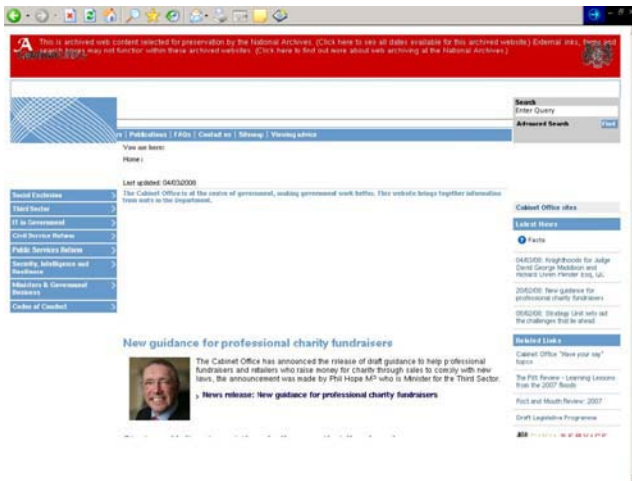


Figure 2. Example of archived webpage where the original layout was adversely impacted by the inclusion of an iFrame.

Currently the use of a server-side include to achieve the desired effect is being developed. Other government website developers and Web publishers solve the problem of the potential dangers of access to inaccurate or out-of-date information by overwriting the information at a given URL as the old information is superseded by new legislation or guidance. While this has an obvious benefit to the organisation and to the user of current information, it causes enormous problems for the user of non-current information – the researcher hoping to compare previous and current policy, the parliamentarian wishing to scrutinize government, the government minister wishing to access documents previously cited by his or her department, the historian wishing to access the record of the past. This issue has proved to be especially problematic in the case of Hansard, which is intended as an historical record of the proceedings of Parliament. When users access a given URL cited in answer to Parliamentary Questions they would often (but perhaps not always) expect to see the information as it was at the time, within the context of the Question that was asked. If information at a URL is regularly overwritten, the answer could be meaningless.

The National Archives has discussed with the Information Services department of the House of Commons the possibility of both parties working on a ‘bridging page’ which would give the user a choice about whether they see historic or current information, but this was considered to be costly, resource intensive, and would require significant redevelopment of both Hansard, its complex publishing routines, and the Government Web Archive.

Both parties consider that the responsibility for the content at URLs must ultimately rest with government publishers and government Web teams. Government organisations do

need to give appropriate consideration to the management of their own content and the user experience.

A further aspect of the temporal expansion of the UK Government Web Archive is the fact that users who unknowingly access the archive through redirection may start to cite the archived URL as a reference in its own right. Consequently, it seemed appropriate to amend the existing contract with the European Archive in order to ensure that a National Archives-specific URL could be developed. The use of a National Archives subdomain, e.g. <http://webarchive.nationalarchives.gov.uk> rather than the previous European Archive-specific URL:

<http://collections.europarchive.org/tna> ensures a strong brand association between the UK Government Web Archive and The National Archives, and clearly identifies The National Archives as custodian of these records.

In terms of the practical implementation of the solution across government, the open source nature of the redirection component in particular, together with the intention to disseminate guidance and links to software downloads has called for a new method of distributing and sharing information. The move away from established IT supplier/department relationships and proprietary software implicit in this project will require government to be more flexible and resourceful in its approach to implementation. The National Archives will be unable to offer a high-level of support to individual organisations because of the resources implicit in such an arrangement. Aside from the technical considerations it has also been recognised that for the new processes, tools and guidance to be effective new groups of stakeholders need to be brought together: central Government website managers, e-communicators, IT staff and those involved in producing web standards for Government. For all these reasons, The National Archives has worked closely with COI, who, after identifying that there are many people working in Web-related fields in government and the wider public sector, has established a collaborative working platform called Digital People. It is intended that the Web Continuity and Sitemaps Implementation Plan sub-communities within this platform will serve as a forum for discussion, support and best-practice sharing for those responsible for website and records management across central government and the wider public sector as a whole. Launched at face-to-face project briefing sessions for central government, run by The National Archives, in May 2008, the forum is intended to complement and utilise existing relationships as well as helping to build new ones.

Within The National Archives, stakeholders from across a number of disparate areas are working together to develop the different elements of the Web Continuity solution: IT, Web, Network, Digital Preservation and Records Management and Cataloguing specialists. The outcome of this work is the potential to bring to the Government web

archive to a much wider raft of stakeholders; most notably academics, wider Government, and most importantly will facilitate better access to Government information for the general public.

The National Archives took forward the solution proposed by the working group in November 2007, and intends that the software, guidance and increased scope of its web archiving programme will be ready by November 2008. The implementation of various elements by Government is expected to take longer, although a system of monitoring will be in place by November 2008, serving as a means of encouraging take-up within government.

Conclusions

The way Government uses the Web has brought many benefits but has also posed questions about long-term access to important information. As a result innovative approaches to Web resource preservation have been required. Following research into the nature of the issues facing users of government information, the working group sought to provide a solution which was both user- and Web-centric. The approach which has been developed draws on The National Archives experience of selective archiving, as well as its expertise in the live Web arena.

The greater number of websites to be archived, as well as the need for the content capture to be as comprehensive as possible, led to the development of a registry database, an automated crawling process and the use of XML Sitemaps. The requirement not only to address the problem of disappearing documents from websites, but the issue of broken links and the implications for the user experience led to the development of the redirection concept. The open source nature of the redirection software, and the bringing together of a wide variety of government stakeholders has made it appropriate for government to harness new social networking tools in order to facilitate discussion and collaborative working. The project has also brought together different groups of stakeholders both within and without The National Archives.

The National Archives has been developing National Collection Strategies to address ellipses in archiving and preservation on a UK-wide basis, encompassing a number of themes and formats, and including Websites. The expansion of the scope of The National Archives Web archiving programme to include Websites of all Central Government departments, agencies and Non-Departmental Public Bodies (NDPBs) positions The National Archives as the source of archived Central Government Websites. In order for the concept of Web Continuity to be truly comprehensive across the UK, The National Archives has been involved in discussions with organisations responsible for the preservation of information pertaining to the devolved administrations of Wales, Scotland and Northern Ireland. The project

has also renewed discussions with other organisations in the Web archiving field, such as the British Library and the members of the UK Web Archiving Consortium on the subject scope and collecting remit of the respective organisations.

Redirection to the Government Web Archive has introduced a temporal dimension to the Web, raising important user considerations, which needed to be addressed through the careful labelling of archived material. Redirection will also bring enormous benefits to the user of the Web, with its potential to bring the Web Archive to a more diverse audience.

“What? So What?”

The Next-Generation JHOVE2 Architecture for Format-Aware Characterization

Stephen Abrams*, Sheila Morrissey**, Tom Cramer***

*California Digital Library
University of California
415 20th Street
Oakland, CA 94612, US
Stephen.Abrams@ucop.edu

**Portico
100 Campus Drive
Princeton, NJ 08450, US
Sheila.Morrissey@portico.org

***Stanford University
314 Meyer Library
Stanford, CA 94305, US
tcramer@stanford.edu

Abstract

The JHOVE characterization framework is widely used by international digital library programs and preservation repositories. However, its extensive use over the past four years has revealed a number of limitations imposed by idiosyncrasies of design and implementation. With funding from the Library of Congress under its National Digital Information Infrastructure Preservation Program (NDIIPP), the California Digital Library, Portico, and Stanford University are collaborating on a two year project to develop and deploy a next-generation architecture providing enhanced performance, streamlined APIs, and significant new features. The JHOVE2 project generalizes the concept of format characterization to include identification, validation, feature extraction, and policy-based assessment. The target of this characterization is not a simple digital file, but a (potentially) complex digital object that may be instantiated in multiple files.

Introduction

Digital preservation is the set of intentions, strategies, and activities aimed at ensuring the continuing usability of digital objects over time. However, since digital objects rely on explicit technological mediation in order to be useful, they are inherently fragile with respect to technological change. Over any significant time period, a gap inevitably arises in the ability of a digital object to function in contemporaneous technological contexts. Put most simply, digital preservation is concerned with effectively managing the consequences of this gap, which is achievable only to the extent to which the gap is quantifiable. The necessary quantification comes, in part, from characterization.

Characterization exposes the significant properties of a digital object and provides a stable starting point for iterative preservation planning and action, as shown in Figure 1 (Brown 2007). Characterization is particularly pertinent to any significant transformative process. The comparison of an object’s pre- and post-transformation properties is a valuable mechanism for quantifying potential transformative loss. In this scenario, the characterization data functions as a canonical

representation or surrogate for the object itself (Lynch 1999).

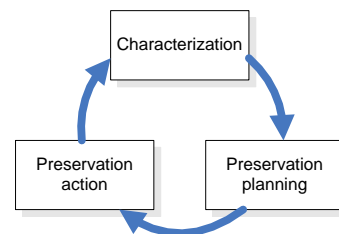


Figure 1. Iterative preservation cycle, adapted from (Brown 2007).

While manual characterization is possible, it is tedious and error prone and requires highly trained staff. Preservation characterization can only be effective at scale through automated efforts (Green and Awre 2007). The original JHOVE framework was developed to provide comprehensive characterization functionality for use in automated systems and workflows (Abrams 2003).

JHOVE was a collaborative project between the Harvard University Library and the JSTOR Electronic Archiving Initiative (now called Portico) with funding from the Andrew W. Mellon Foundation. (More information is available at <http://hul.harvard.edu/jhove/>.) It has found wide acceptance by the international digital library and preservation communities. However, its extensive use over the past four years has revealed a number of limitations imposed by idiosyncrasies of design and implementation. With funding from the Library of Congress under its National Digital Information Infrastructure Preservation Program (NDIIPP), the California Digital Library, Portico, and Stanford University are collaborating on a two year project to develop and deploy JHOVE2, a next-generation architecture providing enhanced performance, streamlined APIs, and significant new features.

Characterization

The description of the original JHOVE framework used the terms *identification*, *validation*, and *characterization* to denote independent concepts. In the context of the JHOVE2 project there has been a shift in terminology under which *characterization* is now defined generically as the totality of description about a formatted digital object, encompassing four specific aspects:

- *Identification*. Identification is the process of determining the presumptive format of a digital object on the basis of suggestive extrinsic hints (for example, an HTTP Content-type header) and intrinsic signatures, both internal (a magic number) and external (a file extension). Ideally, format identification should be reported in terms of a level of confidence.
- *Validation*. Validation is the process of determining a digital object's level of conformance to the requirements of its presumptive format. These requirements are expressed by the normative syntactic and semantic rules of that format's authoritative specification.

Ideally, the determination of conformance should be based on commonly accepted objective criteria. However, many format specifications – particularly those not created as part of explicit standardization efforts – suffer from ambiguous language requiring subjective interpretation. The incorporation of such interpretative decisions into automated systems should be highly configurable to support local variation of preservation policy and practice.

- *Feature extraction*. Feature extraction is the process of reporting the intrinsic properties of a digital object significant to preservation planning and action. These features can function in many contexts as a surrogate for the object itself for purposes of evaluation and decision making.

Note that since digital preservation is concerned with planning for future activities, potentially in response to unforeseeable circumstances, predicting which properties will one day be significant can be problematic. Prudence therefore suggests reporting the most inclusive set of properties possible, while providing sufficiently fine granularity of control to allow for appropriate localized configuration.

- *Assessment*. Assessment is the process of determining the level of acceptability of a digital object for a specific use on the basis of locally-defined policies. Assessments can be used to select appropriate processing actions. In a repository ingest workflow, for example, the range of possible actions could include rejection, normalization, or acceptance in original form.

Reduced to simpler terms, characterization answers the following questions relevant to the preservation of a digital object:

- What is it?
- What is it really?
- What are its salient characteristics?
- What should be done with it?

Or even more reductively, What? and So what?

The JHOVE2 Project

The high-level goals of the JHOVE2 project are three-fold:

- To *refactor* the existing JHOVE architecture and APIs to increase performance, simplify integration, and encourage third-party maintenance and development.
- To provide significant *enhancements* to existing JHOVE functionality to increase its utility to preservation practitioners and workflows.
- To develop JHOVE2 *modules* supporting characterization of a variety of digital formats commonly used to represent audio, geospatial, image, and textual content.

Redesign and Implementation

While JHOVE was implemented in Java 1.4, it used the older stream-style I/O of the standard *java.io* package. JHOVE2 will use the buffer-based NIO package, which has the potential for significantly higher performance through the use of memory mapped I/O (Hitchens 2002).

Although all JHOVE modules implement the same Module interface, and thus share a common method signature, their internal coding is not always similar. Understanding the construction details of one module is not necessarily helpful in understanding the internals of any other module. In order to provide a greater level of conceptual and practical uniformity of implementation, the JHOVE2 design process will establish common design patterns to which all modules will adhere (Fowler 2006). These patterns will also facilitate the integration of individual modules into other systems independent of the core JHOVE2 framework.

The intention of the JHOVE2 project is to continue to provide all existing JHOVE functionality – although implemented in the context of the new framework and APIs – while adding a number of significant new features. The new JHOVE2 code base will be released under the BSD open source license.

More Sophisticated Data Model

JHOVE was designed and implemented with the implicit assumption that a single digital object was equivalent to a single digital file in a single format:

1 object = 1 file = 1 format

(While not strictly true of all modules, the few exceptions to this assumption were dealt with idiosyncratically.) There are, of course, many important

examples for which this assumption is not true. For example, a TIFF file encapsulating an ICC color profile and XMP metadata. While still a single object and file, there are essentially three formats (TIFF, ICC, and XML/RDF):

1 object = 1 file = 3 formats

The JPEG 2000 JPX profile defines a fragmentation feature in which an encoded image can be manifest in an arbitrary number of individual files:

1 object = n files = 1 format

The ESRI Shapefile constitutes a single object that is always manifested by three files, each with its own format:

1 object = 3 files = 3 formats

JHOVE2 data modeling will support the general case of an object manifested by an arbitrary number of component files and formats:

1 object = n files = m formats

From another perspective, however, these kinds of multi-file aggregates can be considered to constitute high-level formats in their own right. For purposes of the JHOVE2 project *format* is defined expansively as a class of objects sharing a common set of syntactic and semantic rules for mapping from abstract information content to serialized bit streams (Abrams 2007). Thus, a page-turning format could be defined consisting of METS descriptive and structural metadata, TIFF master and JPEG delivery page images, and OCR text files:

1 object = 1 + $4n$ files = 1 format

Conceptually, there is no meaningful difference between the traversal of a nested container file – for example, the TIFF with embedded profile and metadata described previously – and a multi-file, multi- directory file system hierarchy. A JHOVE2 module could be developed that would start its recursive parsing at the root “page-turning format” level. As the traversal encounters each lower-level component (image files, OCR files, etc.), JHOVE2 would automatically invoke the appropriate format-specific parser.

In order to support the new concept of arbitrary recursive parsing of complex object formats, three types of identification are needed:

- Identification of the format of *files* based on internal and external signatures.
- Identification of the format of *bit streams* – proper subsets of files – based on internal signatures.
- Identification of the format of *objects* instantiated in multiple files – in other words, a PREMIS representation – based on signatures defined in terms of file-level characteristics and structural relationships.

For example, a Shapefile object can be presumptively identified whenever three sibling files – that is, existing within the same directory

– share a common filename stem but have the extensions *dbf*, *shp*, and *shx*, respectively:

```
abcd/
  1234.dbf
  1234.shp
  1234.spx
```

While object- and file-level identification can occur independent of the parsing necessary for validation and feature extraction, bit stream identification will occur only during the parsing stage.

Generic Plug-in Mechanism

All JHOVE plug-in modules perform the same function – validation and feature extraction – and only a single module is invoked against each digital object. JHOVE2 will implement a more generic processing model in which a configurable sequence of modules, each capable of performing an arbitrary function, can be invoked against each object (see Figure 2). A persistent memory structure for representation information, as defined by the OASIS reference model, will be passed between modules (ISO 2003). Since a given module in the sequence will have access to the results of all subsequent modules, it will be possible to define sophisticated stateful processing flows.

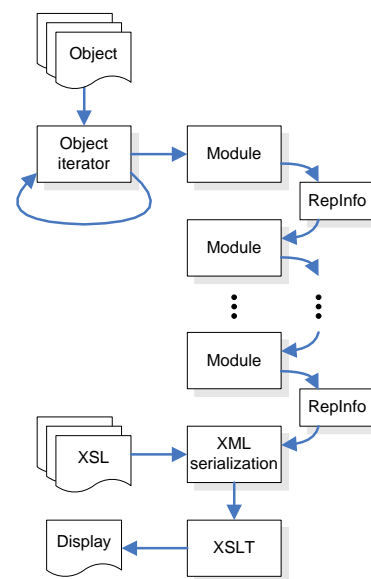


Figure 2. Processing flow.

De-Coupling Identification from Validation

JHOVE performs identification of a digital object’s format by iteratively invoking all configured modules until one reports the object to be valid. Since JHOVE validation is rigorous, this makes identification extremely reliable. However, this benefit is outweighed by the fact that *any* validation error, no matter how trivial, will cause JHOVE to iterate to the subsequent module. Thus, JHOVE will identify a damaged object as, say, a valid bytestream rather than an invalid PDF, which, while technically correct – by definition, *all* objects are valid bytestreams – is not particularly useful in most preservation contexts.

JHOVE2 will de-couple the identification and validation operations. Identification will be performed on the basis of matching file-level characteristics and internal and external signatures. The working assumption is that DROID will be used for file- and bit stream-level identification (Brown 2006).

Standardized Profile and Error Handling

JHOVE modules exist at the granularity of format *families*, but can recognize and distinguish between the many variant formats, or profiles, of the family. For example, the TIFF (Tagged Image File Format) family encompasses a number of specific profiles possessing differences significant in many preservation contexts, such as TIFF/EP, TIFF/IT, GeoTIFF, EXIF, DNG, etc. While at a functional level JHOVE modules provide equivalent handling of profiles, each module's implementation of this function is somewhat idiosyncratic. JHOVE2 will introduce standardized patterns of module design for dealing with profiles in a common and easily extended manner.

Module error handling in JHOVE is similarly idiosyncratic. Again, JHOVE2 will introduce a standardized pattern of error handling with more precise error messages using terminology and references drawn from the appropriate specification documents.

Customizable Reporting

JHOVE is distributed with two output handlers: a Text handler that formats output in terms of simple mail or HTTP header-like name/value pairs, and an XML handler that produces output in terms of a JHOVE-specific container schema. JHOVE2, on the other hand, will always produce an intermediate XML output using a standard METS container schema, which can then be customized through XSL stylesheet transformations to any desired form (Cundiff 2004; Clark 1999). The METS *<StructMap>* mechanism will be particularly useful to model the arbitrary parent-child and sibling structural relationships permitted by the new JHOVE2 object modeling.

The JHOVE2 distribution will include standard stylesheets generating JHOVE-style Text and XML output so that JHOVE2 can easily replace JHOVE in existing workflows dependent upon the specific output form. As with JHOVE, JHOVE2 will report format-specific properties and other important representation information using well-known public schemas such as NISO Z39.87 for raster still images and the forthcoming AES-X098B for audio content (NISO 2006; AES 2008). In addition, the PREMIS schemas will be used for reporting event information and other general preservation metadata (Guenther and Xie 2007).

Modules

Like its predecessor, JHOVE2 will be based on an extensible plug-in framework. Since it is hoped that module development will also occur outside of the context of the JHOVE2 project it is important that JHOVE2 is based on a flexible and robust platform for module integration. The JHOVE2 project will explore the use of the OSGi (Open Services Gateway initiative) and Spring frameworks for this purpose. OSGi provides robust facilities for Java class loading and life cycle

management particularly pertinent for integrating components produced in a decentralized environment (OSGi Alliance 2007). The Spring framework provides a number of functions again useful for simplifying the integration and configuration of disparate components based on the Inversion of Control (IoC) or Dependency Injection paradigm (Johnson et al. 2008).

Module function will include signature-based identification, validation, feature extraction, and assessment. JHOVE2 will also support the humanly-readable display in symbolic form of the contents of binary formatted objects. In JHOVE this functionality was provided in the form of stand-alone utility applications, *j2dump* (for JPEG 2000), *tdump* (for TIFF), etc. In JHOVE2 these functions will be incorporated into the main body of the code. Other function includes API-level support for editing and serializing formatted objects, useful for example to correct existing internal metadata or to embed additional metadata in a syntactically correct manner. It is important to note, however, that an out-of-the-box object editing capability is *not* a project deliverable. JHOVE2 will be an enabling technology for the subsequent development of a number of added-value systems and services, but the development of such products is outside the scope of currently funded JHOVE2 activities.

JHOVE2 will introduce a standard design pattern or template for plug-in modules. This will be based on the "natural" conceptual structures of a given format and their constituent attributes. Each such structure will be mapped to a Java class with methods for parsing, validating, reporting, and serializing; each such attribute will be mapped to a class instance field with appropriate accessor and mutator methods. For example, the major conceptual structures for the TIFF format are the *Image File Header* (IFH) and *Image File Directory* (IFD); for JPEG 2000, the structure is the *Box*; for PDF, the object types *boolean*, *number*, *string*, *name*, *array*, *dictionary*, and *stream*.

Compatibility

As discussed previously, JHOVE2 modules will replicate and extend existing JHOVE functionality. However, due to the nature of the newly proposed features it may not be possible to maintain backwards compatibility with existing JHOVE modules. Compatibility of output will be maintained, however, to the fullest extent possible.

JHOVE2 format identification will be possible for all formats known to the identification module. Presuming the use of DROID, this includes some 580 formats currently documented in the PRONOM database; if the signature database is extended to include the Unix magic number database (*etc/magic*, the basis for the *file* command shell utility), the scope of identification can be extended to over 1000 formats. Detailed validation and feature extraction, on the other hand, is only available for formats for which there are explicit JHOVE2 validation/feature extraction modules.

The JHOVE2 project will provide modules for new formats not supported by JHOVE, including ICC profile, SGML, and Shapefile (ICC 2004; ISO 1986; ESRI 1998). However, budgetary constraints will not permit the reimplementing of all formerly-supported formats;

in particular, modules for AIFF, GIF, HTML, and JPEG are *not* included among project deliverables. It is hoped that subsequent funded activity by project partners or other institutions will quickly remedy these omissions. The remaining JHOVE-supported formats – ASCII, JPEG 2000, PDF, TIFF, UTF-8, WAVE, and XML – will be supported in JHOVE2.

Assessment

One major new function introduced in JHOVE2 is digital object assessment based on locally-defined rules and heuristics. Risk assessment lies at the heart of the preservation decision making process: How can one determine whether a given digital object is approaching incipient obsolescence? What are the factors that make an object susceptible to loss and how can they be quantified? How can an object be evaluated for acceptability under local policy rules? JHOVE2 assessment will be performed by the evaluation of locally-defined rules in the context of prior characterization information. Assessment decisions can be used, for example, to assign appropriate repository service levels, or as factors driving business rules engines to trigger preservation events such as migration (Ferreira, Baptista, and Ramalho 2007; LeFurgy 2002; Pearson and Webb 2007).

The quantitative data necessary to perform such analyses are provided by prior JHOVE2 characterization. Assessment can therefore be seen as the next logical step in a JHOVE2 processing chain:

Identification → Validation → Feature
Extraction → Assessment → Disposition → ...

The JHOVE2 project will investigate existing assessment methodologies and rules, and the means by which they can be codified into best practices and expressed in a highly-configurable, machine-actionable manner (Anderson et al. 2005; Arms and Fleischhauer 2005; Stanescu 2005; van Wijk and Rog 2007).

Schedule

The JHOVE2 project will run for two years. Broadly speaking, the schedule will proceed through three phases:

- Consultation and design (6 months)
- Core framework and APIs (6 months)
- Module development (12 months)

To facilitate communication with and review by important stakeholder communities, the JHOVE2 project will empanel an Advisory Board recruited from leading international preservation institutions, programs, and vendors. Board members will be asked to serve in three capacities: as representatives of the needs of their respective organizations; as proxies for the wider cultural and scientific memory communities; and as independent professional experts.

The capabilities of JHOVE2 described in this paper represent the intentions and plans of the project team at the time of writing. These may evolve, especially during the initial stakeholder consultation period, in order to better serve the needs of the JHOVE2 user community.

More information about the JHOVE2 is available at the project wiki, <http://confluence.ucop.edu/display/JHOVE2Info/Home>.

Conclusion

An understanding of format is fundamental to the long-term preservation of digital objects. While it is possible to preserve digital objects as opaque bit streams without consideration of their format, the end result is merely well preserved bits. In order to recover the information content encoded into those bits requires knowledge of the syntactic and semantic rules governing that encoding, in other words, their format (see Figure 3).


<pre>ffd8ffe000104a46 4946000102010083 00830000ffed0fb0 50686f746f73686f 7020332e30003842 494d03e90a507269 6e7420496e666f00 0000007800000000 0048004800000000 02f40240ffeeffee 0306025203470...</pre>	<pre>SOI APP0 JFIF 1.2 APP13 IPTC APP2 ICC DQT SOF0 183x512 DRI DHT SOS ECS0 ...</pre>	
Syntax	Semantics	Content

Figure 3. Format-directed mapping from JPEG bit stream to humanly-interpretable image content (Burne-Jones 1870-1876). The example image is copyright by the President and Fellows of College Harvard.

The operations of object identification, feature validation, extraction, and assessment lie at the heart of many digital preservation activities, such as submission, ingest (see Figure 4), monitoring, and migration (Figure 5). JHOVE2 will provide a highly configurable, extensible, and functional framework for performing these important operations. Note that Figure 4 shows the deployment of characterization function on both the client *and* server sides of the ingest workflow. The use of JHOVE2 as far upstream as possible in the content lifecycle increases the overall efficiency of preservation activities by facilitating the initial creation of born-preservation amenable content.

JHOVE2 will provide performance improvements and significant new features, most notably, a flexible rules-based assessment capability. The parsing of digital objects underlying JHOVE2 operations will be capable of a recursive traversal of file systems and arbitrarily nested bit streams within files. The revised core framework and APIs will facilitate third-party development and simply the integration of JHOVE2 characterization functionality into existing systems, services, and workflows. The more that JHOVE2 functionality can be dispersed into other open source products and mainstream applications, the more it will benefit from a broader community of use and support.

The JHOVE characterization system has been widely adopted by the international digital memory community.

A number of lessons have emerged from the feedback received from this community. Most significantly, it is now clear that characterization plays a fundamental role in preservation workflows. The JHOVE2 team is very excited to have the opportunity to build on the rich body of prior experience and solidify the foundations for future digital preservation efforts. Through the active input and participation of its stakeholder community, JHOVE2 will remain a central and viable component of preservation infrastructure.

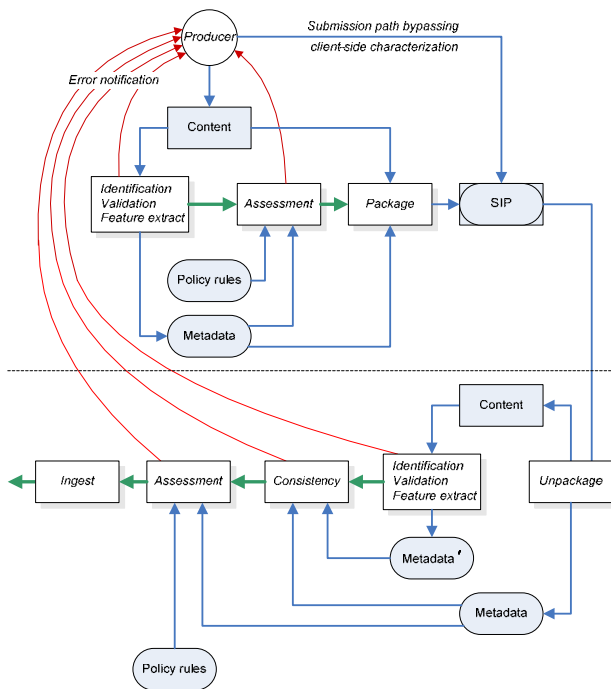


Figure 4. Generic ingest workflow incorporating characterization, adapted from (Abrams 2007).

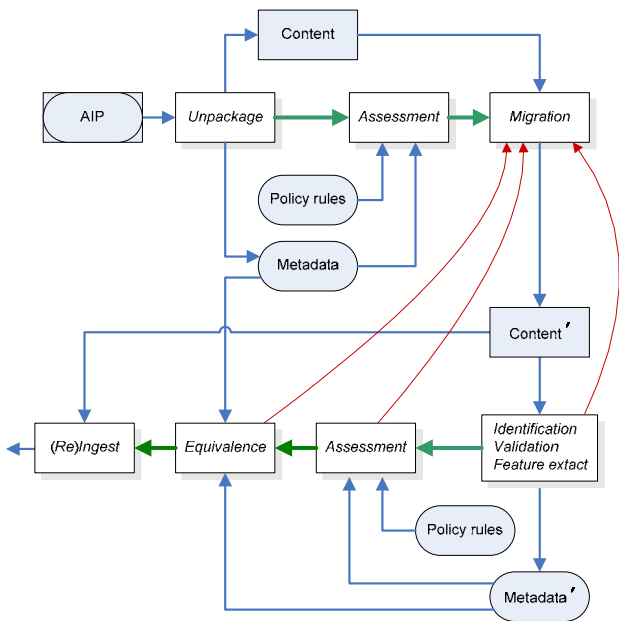


Figure 5. Generic migration workflow incorporating characterization.

Acknowledgements

The JHOVE2 project is funded by the Library of Congress as part of its National Digital Information Infrastructure Preservation Program (NDIIPP). The authors would like to acknowledge the contributions of Justin Littman, Library of Congress program officer, and the JHOVE2 project team: Patricia Cruse, John Kunze, Joan Starr, Hunter Stern, and Marisa Strong at the California Digital Library; John Meyer and Evan Owens at Portico; and Richard Anderson, Hannah Frost, Walter Henry, Nancy Hoebelheinrich, and Keith Johnson at Stanford University.

References

Abrams, S. 2003. Digital Object Format Validation. *Digital Library Federation Fall Forum*, Albuquerque, November 17-19.

Abrams, S. 2007. File Formats. *DCC Digital Curation Manual*.

AES. 2008. Report of the SC-03-06 Working Group on Digital Library and Archive Systems of the SC-03 Subcommittee on the Preservation and Restoration of Audio Recording Meeting.

Anderson, R., Frost, H., Hoebelheinrich, N., and Johnson, K. 2005. The AIHT at Stanford University: Automated Preservation Assessment of Heterogeneous Digital Collections. *D-Lib Magazine* 11(2), December.

Arms, C., and Fleischhauer, C. 2005. Digital Formats: Factors for Sustainability, Quality, and Functionality. *IS&T Archiving Conference*.

Brown, A. 2006. Automated Format Identification Using PRONOM and DROID. Technical Paper DPTP-1, Issue 2, March 7.

Brown, A. 2007. Developing Practical Approaches to Active Preservation. *International Journal of Digital Curation* 2(1): 3-11.

Burne-Jones, E. 1870-1876. The Days of Creation: The First Day. Harvard University Art Museums, 1943.454. JPEG image, http://via.lib.harvard.edu/via/deliver/deepLinkItem?recordId=HUAM303460&componentId=HUAM:51430_mddl.

Clark, J., ed. 1999. XSL Transformations (XSLT). Version 1.0, W3C Recommendation, November 16.

Cundiff, M. 2004. An Introduction to the Metadata Encoding and Transmission Standard (METS). *Library Hi Tech* 22(1): 52-64.

ESRI. 1998. ESRI Shapefile Technical Description. July.

Ferreira, N., Baptista, A., and Ramalho, J. 2007. An Intelligent Decision Support System for Digital Preservation. *International Journal on Digital Libraries* 6(4): 295-304.

Fowler, M. 2006. Writing Software Patterns. Web site, www.martinfowler.com/articles/writingPatterns.html, accessed August 9, 2008.

Green, R., and Awre, C. 2007. RepoMMAN Project: Automatic Generation of Object Metadata, Technical Report D-D13, Version 1.1, October.

Guenther, R., and Xie, Z. 2007. Implementing PREMIS in Container Formats. *IS&T Archiving Conference*.

Hitchens, R. 2002. *Java NIO*. Sebastopol: O'Reilly.

ICC. 1:2004-10. 2004. Image technology colour management – Architecture, profile format, and data structure. Version 4.2.0.0.

ISO 14721. 2003. Space data and information transfer systems – Open archival information system – Reference model.

ISO 8879. 1986. Information processing – Text and office systems – Standard Generalized Markup Language (SGML).

Johnson, R., et al. 2008. The Spring Framework – Reference Documentation.

LeFurgy, W. 2002. Levels of Service for Digital Repositories. *D-Lib Magazine* 8(2), May.

Lynch, C. 1999. Canonicalization: A Fundamental Tool to Facilitate Preservation and Management of Digital Information. *D-Lib Magazine* 5(9), September.

NISO Z39.87. 2006. Data Dictionary – Technical Metadata for Digital Still Images.

OSGi Alliance. 2007. About the OSGi Service Platform. Technical Whitepaper, Revision 4.1, June 7.

Pearson, D., and Webb, C. 2007. Defining File Format Obsolescence: A Risky Journey. *3rd International Digital Curation Conference*. Washington, December 11-13.

Stanescu, A. 2005. Assessing the Durability of Formats in a Digital Preservation Environment: The INFORM Methodology. *OCLC Systems & Services* 21(1): 61-81.

van Wijk, C., and Rog, J. 2007. Evaluating File Formats for Long-term Preservation. *4th International Conference on Preservation of Digital Objects*. Beijing, October 11-12.

Emulation: From Digital Artefact to Remotely Rendered Environments

Dirk von Suchodoletz, Jeffrey van der Hoeven

University of Freiburg, Koninklijke Bibliotheek

Fahnenbergplatz, 79085 Freiburg, Germany

Prins Willem-Alexanderhof 5, 2509 LK Den Haag, The Netherlands

dsuchod@uni-freiburg.de, jeffrey.vanderhoeven@kb.nl

Abstract

Emulation used as a long-term preservation strategy offers the possibility to keep digital objects in their original condition and experience them within their original computer environment. However, having only an emulator in place is not enough. To apply emulation as a fully-fledged strategy, an automated and user-friendly approach is required. This can not be done without knowledge of the original software and contextual information about it.

This paper combines the existing concept of a *view path*, which captures contextual information of software, together with new insights to improve the concept with extra metadata. It provides regularly updated instructions for archival management to preserve and access its artifacts. The view path model requires extensions of the metadata set of primary object of interest and depends on additionally stored secondary objects for environment recreation like applications or operating systems.

This paper also addresses a strategy to render digital objects by running emulation processes on distance. The advantages of this strategy are that it improves user convenience while maximizing emulation capabilities.

Challenges in Long-term Preservation

Unlike books, newspapers, photographs or other traditional material, digital objects require a digital context consisting of a combination of software and hardware components. Due to technological advance hardware and software becomes obsolete leaving it uncertain if we still can render today's digital objects in the future. Permanent access to archived digital artifacts thus raises challenges to archive operators who have to deal with keeping access to digital material without loss of information.

Several solutions for long-term access exist of which migration and emulation are the main flavors. Migration – the mostly used digital archiving strategy today – tries to address this problem by changing the digital object to prepare it for access and rendering in future digital

environments. Although this strategy is applicable for static digital objects such as images, text, sound and animation, it is not suitable for dynamic objects such as educational software or computer games. As a lot of digital material is becoming more advanced, solely relying on migration as preservation strategy is risky and will certainly result in loss of authenticity and information.

Emulation offers a different approach. It does not change the digital object itself, but tries to recreate the original computer environment in which the object used to be rendered. Each layer of the software-hardware-stack can be used as a working point for emulation: applications, operating systems or hardware can be recreated in software by using an emulator for the actual environment.

However, an emulator is relying on a computer environment as well. From the perspective of archive management emulators do not differ significantly from other digital objects. Even emulators become obsolete with the evolution of digital environments. Several strategies exist to keep the emulators available in a changing environment (Verdegem and Van der Hoeven 2006). For example, the Koninklijke Bibliotheek (KB) and Nationaal Archief of the Netherlands developed Dioscuri (Dioscuri 2008), an x86 emulator developed with the purpose for long-term archiving kept in mind (Van der Hoeven, Lohman, Verdegem 2008). This emulator bridges the widening gap between older x86 machinery and recent architectures by using a virtual layer between operating system and emulator to abstract from specific reference platforms. Furthermore, detailed documentation on every step taken in design and development is being preserved to allow future users and developers understand the software.

Bridging the Past to the Future

No matter which emulator is chosen, contextual information of the computer environment is always required. For example, questions such as "for which operating systems is WordPerfect 5.1 compatible?" are less obvious today than twenty years ago. To overcome this gap of missing knowledge, a formalization process is needed to compute the actual needs for an authentic rendering environment of the digital artefact. In 2002, IBM Netherlands proposed the concept of a *view path* based on

Copyright © 2008, the authors.

Work presented in this paper is partially supported by European Community under the Information Society Technologies (IST) Programme of the 6th FP for RTD - Project IST-033789. The authors are solely responsible for the content of this paper. It does not represent the opinion of the European Community, and the European Community is not responsible for any use that might be made of data appearing therein.

their *Preservation Layer Model (PLM)* (van Diessen 2002) which has been refined during the research on emulation at Freiburg University and the European project Planetes.

The PLM outlines how a file format or collection of similar objects depends on its environment. A PLM consists of one or more layers of which each layer represents a specific dependency. The most common PLM consists of three layers: application layer, operating system layer and hardware layer. However, other variations can be created as well. Based on a PLM, different software and hardware combinations can be created. Each such combination is called a view path. In other words, a view path is a virtual line of action starting from the file format of a digital object and linking this information to a description of required software and hardware. Figure 1 illustrates some typical view paths starting for a particular digital object. Depending on the type of object a specific rendering application is required. This application requires a certain operating system (OS) to be executed whereas in turn the OS relies on particular hardware.

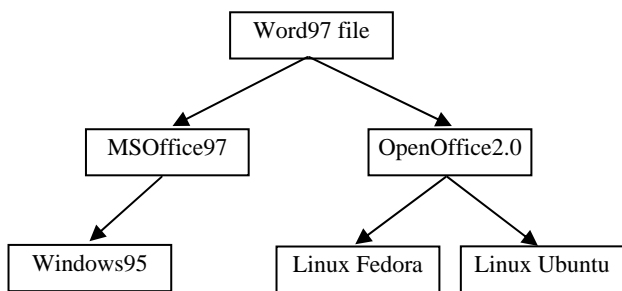


Figure 1: example view paths

As dependencies might change in the future, once derived view paths may change over time as well. This is the case when certain hardware and software become obsolete. To solve this missing link the dependency can be replaced by another compatible environment or by using emulators to bridge the gap between the digital past and future (figure 2).

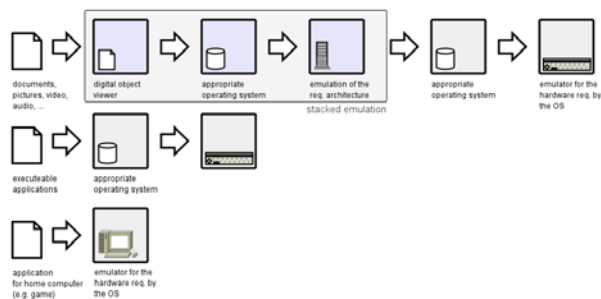


Figure 2: emulation and view paths

Looking into the future, the following situations regarding object dependencies can occur:

- At a given point in time there is exactly one view path for an object to its rendition.
- An object has become inaccessible because all view paths have become obsolete.
- There are several different view paths for a digital object available which require a selection procedure.

The first situation leaves no discussion as there is only one way to retain access to the digital object. The second situation needs some additional processing. Apparently, one or more layers of the view path have become obsolete. This can be solved by using emulators instead.

The third situation however is not easily decided. To manage various rendering options a procedure will be needed to find the best or most preferred view path for rendering a certain object or collection of objects.

View Path Extensions using Metrics

To apply the PLM in combination with emulation in archival management a formalization and automation of the decision process is required. To do so, the model can be extended with metrics. A metric could be any kind of measurement along a view path and can be created by attaching a certain weight on a subsection of a view path. Current metadata for a layer in a view path needs to be extended to capture metric information. Having applied weightings to all view paths, a classification can be made by what the metric stands for. In general, using view path metrics offer the following possibilities (see also figure 3):

- allow comparison of each option to ensure a high grade of authenticity and quality of the object rendering or execution;
- offer quantifiers to emphasize on particular aspects, such as authenticity or ease of use;
- to include the archive users preferences, in the field of applications, operating systems or reference platforms;
- to allow cost-benefit analysis quantifying which view paths are economically feasible and which are not.

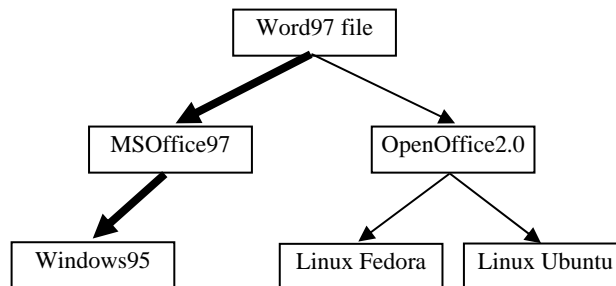


Figure 3: example view paths with weighting denoted by arrows

Assigning weights to a view path based on the authenticity of a certain computer environment, can help a user find the most authentic representation of a digital object. Also, weights could be altered to influence a requested rendering in a certain direction. Furthermore, users of computer environments can help evaluating view paths by adding reviews and ratings to a view path based on quality, completeness and correctness, or ease of use.

Another way in which view path metrics can be helpful is managing costs of preserving the original environment. During the preservation period of each digital object the determined view path for the associated object type has to be checked on every change of the reference environment. For example, obsolete hardware or updates for software affects the object's dependencies and should therefore be considered in the associated view path as well. Furthermore, changing the view path also requires changes in the actual emulation environment resulting in various updates in hardware, software and configurations.

Maintenance of each view path and environment brings certain costs with it. Attaching metrics representing operational costs to each view path can deliver an estimate for the effort to spent to archive a specific type of object. If these costs pass a certain threshold, economic considerations could be taken into account when ingesting the objects into the archive. This knowledge may help to suggest on formats or prefer specific types of objects over others.

If multiple view paths exist for a given object type, costs would offer another metric to decide on good alternatives. For example, assume a digital object is formatted according the PDF 1.0 Standard and is accessible with a tool for MS Windows 3.11. If there are other tools offering the same results in quality and authenticity and there are no other object types requesting this specific view path it might be advisable to drop this environment and aggregate the paths. This not only lowers administration costs for keeping the view path current, but also reduces costs for preserving the necessary software in an archive.

Digital Archive Management

In perspective of emulation, several tasks have to be carried out to ensure that a digital object is preserved and accessible in a long-term digital archive. In general, three phases can be distinguished: the required workflow steps on object ingest, the periodical operational procedures of archive operation and the procedures for the object digest to the interested user of a digital object.

On ingest, identification and characterisation of digital objects have to be performed. Several solutions exist already of which the most prominent at the moment are PRONOM and DROID of The National Archives in the UK (The National Archives, 2008). However, as these

tools are able to offer information about the digital format and some of its dependencies, they do not take into account all computer related dependencies such as hardware and emulators. Therefore, extensions should be made to incorporate the PLM and its extensions for keeping track of metrics in the model.

Another important part of archive management is the selection of proper emulators. At ingest time and during the whole period of preservation, availability of view paths have to be checked and if a view path has become obsolete, emulators can be used to close the gap between the layers in the path. If no suitable view path can be constructed, the digital object might be rejected at ingest time because no guarantee can be given that it will remain accessible over the long term.

Having an emulator and contextual information contained in a view path still leaves some implications at the time a digital object is disseminated and needs to be rendered. Firstly, the original environment consisting of software and hardware needs to be preserved. Secondly, an emulation service is needed to reconstruct the original environment, configure the emulator and activate the emulation process. In the next sections, these two topics will be discussed in more detail.

Software Archive

In recent years, a lot of attention has been paid to emulation and virtualisation software as the primary requirement for retaining access to any kind of digital information authentically. However, emulators only solve one part of the equation. Additional software such as operating system and applications are needed as well (Reichherzer and Brown 2006).

Currently, no standardized or coordinated approach for software preservation exists. Some national libraries treat software releases the same way as publications and preserve them on the shelf next to their books and journals (BnF 2008). Although these software are indexed and managed, the actual bits are still on their original media carriers and are not directly accessible for library visitors. Media deterioration is a serious threat and might result in loss of information in the near future. Other organisations, such as the KB or University of Freiburg rely on external sources such as software companies to take care of preserving released software.

To better understand why this area is not yet covered, several reasons can be given that obstruct preservation and access to software. Firstly, the newer the computer environment is, the higher the level of complexity and the number of additional software components needed. Current computer systems are running very complex applications which rely on a wide range of utilities such as hardware drivers, plug-ins, video and audio decoders, fonts and

many more. Preserving such an application implicitly means all depending sources need to be preserved as well.

Secondly, legal issues arise. Digital rights management and copy protection mechanisms can prevent one from copying the original bit stream from its carrier into a digital archive. Even if it is technically possible to preserve it, legal implications still exist that might forbid future generations to use the software. To complicate matters, some software require an online activation or regularly updates to remain operational. Having the software package itself does not work for future usage.

A third issue is to understand how software operates. This might be obvious today, but can become problematic in the future. Extensive metadata is required to cope this problem, addressing not only the title and release date of software but also more semantic information such as installation manuals, tutorials and reference guides.

An final interesting challenge is the diversity of software releases. Most software is adapted to different human languages, geographical areas and units of different parameters. The latter include various currencies, their representation with some specific characters, dimensions or the sizes, format of date, calculations and the number of public or religious holidays.

Aside from these obstructions, keeping digital objects alive via emulation the original software needs to be preserved as well. For safekeeping emulators, operating systems, applications and utilities similar guidelines as for digital objects can be applied. That is, software should be stored under the same conditions as other digital objects by preserving them in a OAIS-based (ISO 14721:2003) digital archive.

Nevertheless, it might be of interest to retain various access copies of software to allow emulators to prepare them for convenient use during the realization of a view path. Often requested view paths could be stored as combined caches of applications, operating systems and the emulator for faster access. Such specifically prepared containers could be distributed between memory institutions to share the load of management overhead and costs.

Remote access to emulated environments

Assuming that the required software and metadata is available for emulation, the environment to be emulated has to be prepared. This is a very technical process and

requires skilled personnel to merge all required software into one computer environment, set emulator parameters and offer guidance to the user about how to work with ancient computer environments.

To tackle this challenge it would be desirable to centralize the whole process in specialized units with trained personnel and offer services within a framework over internet. This eases the complex procedures to run emulators and reduces the system requirements of the user to a viewer, preferably a web browser. The user gets the results presented via a virtual screen remotely on its computer. In overview, this kind of setup would offer the following benefits:

- Access to digital objects is location independent.
- No special system requirements at the user's side is necessary.
- Management of such a service can be centralized and several memory institutions could share the workload or specialize on certain environments and share their expertise with others.
- Problems of license handling and digital rights management could be avoided, because software does not need to be copied onto users private machinery but instead only runs at the service provider.
- Organisations such as computer museums are able to present their collections in an alternative way as they are not restricted to one room.

Still, knowledge about old computer environments is needed to work with emulated computers, but on-screen instructions might offer an extra aid.

Within the Planets project, a pilot is being carried out by developing a prototype of an emulation service. This service is based on existing emulators such as Dioscuri and allows them to run on a remote basis. Transportation of the remotely rendered environment is done by GRATE which stands for Global Remote Access to Emulation Services and is currently under development by the University of Freiburg. With GRATE any user can easily access emulated environments on distance via their web browser.

First experiments prove that this solution is very user friendly and flexible in configuration. Figure 4 and 5 show two screenshots of GRATE. The first one runs Dioscuri on distance loading WordPerfect 5.1. The second image shows the desktop of Windows98 executed by QEMU emulator (QEMU 2008). Both emulated environments are accessed via a normal web browser.

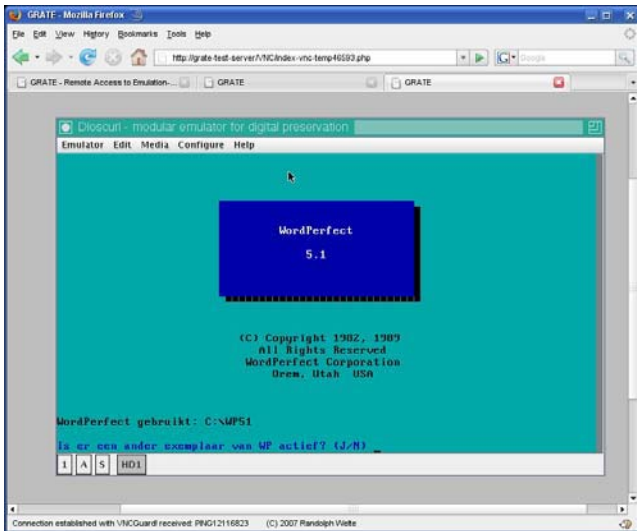


Figure 4: GRATE running Dioscuri with WordPerfect5.1

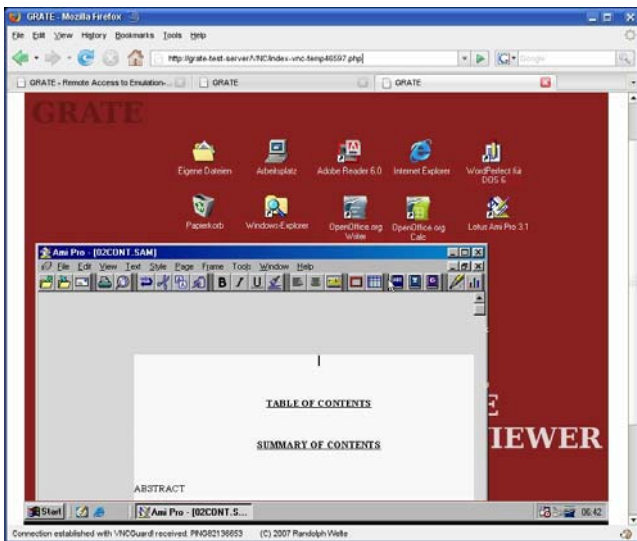


Figure 5: GRATE running QEMU with Windows98

The next step is to integrate this emulation service with the interoperability framework of Planets. This will result in a major extension in functionality for preservation action strategies allowing a Planets-user to automatically start emulation activities when a digital object needs to be rendered in its authentic computer environment.

Conclusions

Emulation strategies help to offer sustainable access to digital objects in their authentic environment. Aside from an emulator, other conditions have to be met for successful rendering of the object. Specific information about the object's dependencies on hardware and software should be preserved. Furthermore, to recreate an old computer

environment, access to the original software is needed and a access mechanism is required for configuring the emulator and environment.

A flexible solution to manage metadata of environmental dependencies is by using the Preservation Layer Model (PLM). The PLM introduces view paths for each combination of hardware, software and digital file format or collection of files. These formalizations can aid the archivists to manage their digital collections and give them guidelines what to do on object ingest, during object storage and on dissemination. However, the current PLM structure does not explain what to do when multiple view paths can be applied. To overcome this, view path metrics could help to optimize this operation by attaching weights to each layer of dependencies. These weights can be influenced by several reasons such as “most common used operating system” or “user preferred application”. The community could even have a vote in the selection procedure by offering feedback and ratings. Moreover, a cost/benefit analysis can be applied to drop less effective view paths.

Aside from metadata the actual software is needed. As software is a crucial piece of the puzzle for emulation, initiatives have to be taken to start preserving software for the long term. This is a task that requires a coordinated action because it is of interest for all organizations that would like to retain authentic access to digital objects.

To simplify access to emulated environments, a remote emulation service is proposed. Currently, both the Koninklijke Bibliotheek and the University of Freiburg are involved in creating such a service based on the emulator Dioscuri and GRATE, a specialized remote emulation transport tool. Further refinement of this approach within the Planets project will result in the next generation of emulation services, offering centralized access to emulated environments via a generic web interface.

References

Verdegem, R., and Van der Hoeven, J.R., Emulation: To be or not to be. In proceedings of the Archiving 2006 Final Program and Proceedings, 55 – 60. Ottawa, Canada: IS&T Archiving Conference.

Dioscuri, the durable x86 emulator in Java, <http://dioscuri.sourceforge.net>

Van der Hoeven, J.R., Lohman, B., Verdegem, R., Requirements for Applying Emulation as a Preservation Strategy. In proceedings of the Archiving 2008 Final Program and Proceedings. Bern, Switzerland: IS&T Archiving Conference.

Van Diessen, R.J., 2002. Preservation Requirements in a Deposit System, Technical Report, IBM / KB Long-term Preservation Study.

PRONOM & DROID, The National Archives (TNA), <http://www.nationalarchives.gov.uk/pronom/>
<http://droid.sourceforge.net/>

Reichherzer, T., Brown, G., Quantifying software requirements for supporting archived office documents using emulation. In proceedings of Proceedings of the 6th ACM/IEEE-CS joint conference on Digital libraries, 86 – 94. Chapel Hill, USA, International Conference on Digital Libraries (ICDL) 2006.

Bibliothèque nationale de France (BnF), <http://www.bnf.fr>

QEMU website, <http://bellard.org/qemu/>

Acknowledgements

Planets is co-funded by the European Union under the Sixth Framework Programme. Planets is a substantial collaborative project that builds on and brings together the work of many talented individuals contributing from a consortium of committed organisations.

Data without meaning: Establishing the significant properties of digital research

Gareth Knight¹ and Maureen Pennock²

1. Centre for e-Research, Kings College London, 26 - 29 Drury Lane, London, WC2B 5RL

2. UKOLN, University of Bath, Bath, BA2 7AY

gareth.knight@kcl.ac.uk; m.pennock@ukoln.ac.uk

Abstract

It is well recognised that the time period in which digital research may remain accessible is likely to be short in comparison to the period in which it will have intellectual value. Although many digital preservation strategies are effective for simple resources, it is not always possible to confirm that all of the significant properties – the characteristics that contribute to the intended meaning – have been maintained when stored in different formats and software environments. The paper outlines methodologies being developed by InterPARES, PLANETS and other projects in the international research community to support the decision-making process and highlights the work of four recent JISC-funded studies to specify the significant properties of vector images, moving images, software and learning objects.

Introduction

In recent years, there has been a growing awareness of the need for digital preservation to maintain access to digital research. Unlike physical artefacts, it is considered to be infeasible to store digital data in its original form and expect it to be readable and usable over time [6]. Instead, there is an expectation that the environment in which digital records are accessed will change on an ongoing basis, e.g. as a result of updates to the computer hardware, operating system, or application software in use [24]. Institutions with a commitment to maintain digital research may adopt several digital preservation strategies, such as format conversion (normalisation, migration), emulation of the original hardware and software and, for certain types of data, re-implementation according to an existing specification. This paper will introduce the concept of significant properties and its role in maintaining the authenticity of research data across changing technological environments over time. It will highlight criteria for the evaluation of significant properties, through consideration of the requirements of those that have an investment in the availability and use of digital research. It will subsequently highlight work that has or is being performed to assist institutions with the task of understanding and evaluating significant properties. A final section provides a comparative analysis of the significant properties of vector images, moving images, software and learning objects that were identified by four recent JISC-funded studies.

Definitions of significant properties

The term ‘significant properties’¹ was first used by the CEDARS Project [5] and has been interpreted using several different, but broadly consistent definitions [7]. For the purpose of this paper, significant properties are defined as the characteristics of an information object that must be maintained to ensure its continued access, use, and meaning over time as it is moved to new technologies [24]. The term is widely used in the archival community, where it is associated with authenticity (that it is what it purports to be) and integrity (that it has not been changed or corrupted in a manner that has caused the original meaning to be lost) [24, 9, 3]. Significant properties share some similarities with Representation Information and there is some crossover between the two concepts. In an OAIS, significant properties are the characteristics of the abstract information object (e.g. an image), while representation information indicates characteristics of the data object (e.g. format, encoding scheme, algorithm) [2].

Research on the topic of significant properties

The importance and position of significant properties in developing digital preservation strategies has been recognised by several parties over the past decade. The following list is not intended to be exhaustive, rather an illustration of the projects that have made an important contribution to the development of our understanding of significant properties:

- *CEDARS (Curl Exemplars in Digital ARchiveS)*: the JISC-funded CEDARS project (1998-2002) explored several digital preservation issues, including significant properties. The project defined the ‘Underlying Abstract Form’, an abstract model for preserving ‘all the necessary properties of the data’ [5].
- *Digital Preservation Testbed*: Complementary research took place in the Dutch Digital Preservation Testbed project (2000 – 2003) testing the viability of

¹ essence, essential characteristics, core features, properties of conceptual objects are other synonyms that are used in particular domains and institutions.

different preservation approaches for different types of government archival digital records. The research was based on the assumption that different types of records have different preservation and authenticity requirements [18]

- *National Archives of Australia*: The NAA developed the concept of the ‘essence’ as a formal mechanism to determine the characteristics that must be preserved and a ‘Performance model’ to demonstrate that digital records are not stable artefacts; instead they are a series of performances that change across time [14].
- *DELOS*: The preservation cluster in the EU-funded DELOS Network of Excellence in Digital Libraries built on the work of the Testbed project and developed a metric for testing and evaluating digital preservation strategies using utility analysis and an Objective Tree [19].
- *PLANETS*: PLANETS is an EU-funded project that is undertaking several projects that have relevance to the description of significant properties, including the continued development and integration of the DELOS Utility Analysis and Objective Tree into the PLATO Preservation Planning Tool and the creation of the eXtensible Characterisation Definition/Extraction Language (XCDL/XCEL) [17].
- *JISC-funded Significant Properties projects*: the JISC has funded four short projects to investigate the significant properties of vector graphics, moving images, learning objects and software that have produced some useful outputs [10].
- *InSPECT Project*: InSPECT is a JISC-funded two-year project performed by the Centre for e-Research at Kings College London and The National Archives. It is building on the work performed by the National Archives of Australia and Digital Preservation Testbed to develop a framework for the definition and description of significant properties, which will be integrated into the PRONOM format registry [12].

Although each project has a distinct conceptual basis and methodology, the outputs of earlier work has contributed to the development of subsequent projects.

Criteria for evaluating significant properties

An implicit assumption in the use of terminology, such as ‘significant’ and ‘essential’ is the recognition that criteria is required against which the relative value of each property may be assessed. The Oxford English Dictionary defines ‘value’ as a noun to be ‘a fair or adequate equivalent or return’. In diplomacy a distinction is made between ‘intrinsic value’ - that something has value ‘in its own right’- and ‘extrinsic value’ - that value is derived from an external function. The InterPARES Authenticity Task Force has hypothesised that both intrinsic and extrinsic elements will play key roles in establishing the

identity of a digital record [15]. For digital objects, value judgments made by an archivist or collection manager will determine the level of functionality that is retained in subsequent iterations of the object. It is therefore important to identify the potential stakeholders and understand the functions that will be required of the information object and the environment in which it will be used, as criteria for evaluating alternative preservation strategies [20, 5, 21]

The InSPECT project [12] has analysed several elements that may influence an institutions interpretation of value and, as a result the preservation activities that must be performed to maintain the various properties of the information object. These may be summarized into four categories:

1. Stakeholder requirements

The stakeholders represent the intended audience for the digital object. The consideration of the required functionality that an Information Object should provide must consider several stakeholders during its lifecycle. These may include:

- 1) The creator who produced the resource to fulfil specific aims and objectives in the short-term. For example, a paper written for publication.
- 2) Researchers in the designated community who wish to use the resource as the basis for further analysis and discussion, e.g. scientists, artists.
- 3) Tutors who wish to incorporate the resource into a learning object for use in teaching [1]

In addition a digital curator should be aware of their own requirements:

- 4) A curatorial institution that wishes to maintain an authentic copy of the resource for the purpose of curation and preservation.

The functionality required by each stakeholder may differ and change over time, influenced by aims and objectives directly defined by the stakeholder or imposed by business requirements (e.g. legal status, basis for funding, mandate, institutional policy of other stakeholders). Although a full analysis is required, it is reasonable to suggest that some or all stakeholders will require the digital object to be authentic. Each stakeholder will have different criteria for evaluating authenticity, which is influenced by the context of their work. For example, the InterPARES project [15] notes that the authenticity requirements for legal records are strict which requires the adoption of a risk-adverse strategy to preservation. In comparison, the authenticity requirements for a funding body may be much lower, limited to the requirement to maintain the intellectual content of the resource only [21]. A second function that may be required is the ability to use and modify content by the creator or a third-party, in addition to the ability to access it. For example, the ability to search and edit a spreadsheet, database, and word processing document have

been cited as potential useful functions that support the activities of financial institutions [19, 21].

2. Type of resource

The method in which a Creator first expresses an idea and renders it in a form that can be understood by others has an influence upon the properties that are considered to be significant. The creation process may be influenced by the design preferences of the Creator (e.g. an idea expressed as a page of text, a spider diagram, or audio recording), the software tools available, as well as consideration of the access method for a target audience. To illustrate the distinction between object types, a report may be written for communication in an email or a word processing document. Both will have common properties that are specific to the form of expression (words organised into paragraphs) and the method of embodiment (e.g. title may be indicated in subject line of an email or document body). However, an email will require additional properties to record details of the recipient.

3. Legal right

The copyright of digital research may be owned by one or more stakeholders. An institution with a commitment to curate and preserve the significant properties of a digital resource may be limited in its actions by the legal rights that have been assigned to it, which will limit the range of properties that it is capable of maintaining. For example, a research paper may contain text and images owned by the author that may be reproduced in a different format and typographical features owned by a publisher that cannot be reproduced [22].

4. Capability

Finally, the ability of the curator to perform preservation action for digital research may be influenced by the total money, time and resources available for the identification and evaluation of properties. The institution may have possess sufficient finances to purchase or develop a software tool to perform a data analysis; to allocate staff time to the identification of significant properties; and/or validate that they have been maintained in subsequent manifestations.

The creation of a definition of significance encompasses a range of qualitative requirements that may be unique to each institution. The PLANETS PLATO tool [17] may prove useful through the provision of a baseline set of characteristics that can be tailored to the requirements of each institution.

Framework for the evaluation of significant properties

The creation of a framework for the identification and analysis of significant properties has been a key area for research in recent years. The work of Rothenberg & Bikson [21], DELOS [19] and the InterPARES projects [15] has been particularly influential in this area. The

following section provides a description of two frameworks, Digital Diplomats and Utility Analysis that may assist curators to interpret the properties of digital research that must be maintained.

Digital Diplomats

Digital diplomats is the application of archival diplomats to digital records, which was developed for use in the InterPARES1 project. The process emerged in the seventeenth century as a method for determining the authenticity of a physical record for legal purposes. On the basis of the examination, it may be possible to establish if the document was created at the time and place that is claimed. In comparison to other methodologies, their analytical method places a greater emphasis on the intended function (e.g. a legal document) that the record must perform as a basis for defining the significant properties. The InterPARES project indicates that many authenticity requirements are created and managed at an organizational level, and therefore cannot be entirely understood at the record-level. To demonstrate the application of diplomats to digital records, they indicate that properties may be organized into four categories:

1. *Documentary form*: The elements that establish its authority in an administrative or documentary context. These are separated into intrinsic and extrinsic elements. Intrinsic elements specify the context in which the record exists. For example, details of the creator, intended recipient, date of creation, and aspects that communicate the activity in which it participates. Extrinsic elements refer to the perceivable features that are instrumental in achieving an intended purpose. For example, the overall presentation of the intellectual content (text, image, sound), presentation features specific to the record (e.g. special layouts, hyperlinks, colours, sample rate), electronic signatures, digital time stamps and other 'special signs' (watermarks, an institution's logo).
2. *Annotations*: The aspects of the record that have been augmented after its creation. For example, additions made as part of: its execution (datetime that an email was transmitted, indication of attachments); its handling in relation to its intended use (comments embedded in the record that critique the work); and its handling for records management purposes (identifier, version number, cross reference to other records).
3. *Context*: the broader framework in which the record is created and managed. For example, judicial-administrative, documentary and technological context.
4. *Medium*: Diplomatic analysis specifies the medium on which information is stored as an essential element. However, the InterPARES Authenticity Task Force indicates that an analysis of the medium is transitory and may be an unnecessary consideration for many digital records.

The classification of different aspects of a digital object is a useful stage in the evaluation of the aspects that should be considered significant, in relation to one or more intended functions. However, the use of archival diplomatics as an analytical tool imposes certain well recognised limitations on the type of information that is considered to be significant. Specifically, there is an emphasis on textual elements of agents associated with the creation, augmentation and management process. The project has also noted the requirements for ‘fixed form’ records, which excludes certain types of dynamic data [15]. The approach taken by the InterPARES1 project in establishing the contextual basis for decisions at an organisational-level is useful, but further work is necessary, potentially based on less strict compliance with archival diplomatics analysis.

Utility Analysis

The preservation cluster in the EU-funded DELOS project built on the work of the Testbed project to develop a metric to test and evaluate digital preservation strategies, based on the conceptual Utility Analysis and Objective tree [19,17]. The metric may be used to define objectives and evaluate the results of preservation activities. The Utility Analysis model specifies eight stages (figure 2)

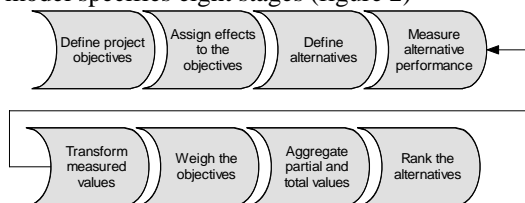


Figure 2: the eight steps of the DELOS Utility model

In the DELOS Utility Analysis and Objective tree, significant properties of digital objects are one of several factors that must be considered when defining and subsequently evaluating objectives. They may be divided into two major groups: ‘file characteristics’ that indicate the aspects of the digital object that must be maintained (e.g. horizontal and vertical dimensions of an image, frame rate of moving image) and ‘process characteristics’ that describes the objectives with which the resulting digital object must comply (e.g. authentic recreation of the significant properties, scalability, error-detection, usability, and others). The metrics developed in DELOS may be used to automatically weigh the performance of a given approach in preserving specific characteristics of records and the numerical evaluation of preservation strategies is consider to be a step towards the automation of the evaluation process.

To demonstrate their approach the project carried out two case studies [19], indicating the requirements of a word processing document and an audio file. The analysis of the file characteristics in a word processing document

identified a number of properties that must be maintained, including various aspects of the content (body text, embedded images, foot notes, page numbering), page layout (paragraphs, page margins, page breaks) and function of the creating application (Microsoft Word). The latter is surprising, but is supported by earlier work by Rothenberg & Bikson [21]. In terms of the process characteristics, the ability to track changes and search the document was considered to be significant. The criteria was subsequently used as a basis for evaluation of suitable file formats, indicating that the most suitable format to contain the ‘file characteristics’ and ‘process characteristics’ was another version of Microsoft Word. Whilst the high score may be due to fundamentally necessary compatibility between the source and target file formats, some would consider this an undesirable route in terms of format longevity. It is clear that any attribution of measured value can be subjective and is not necessarily transferable to other situations; different organisations with different baseline requirements will likely allocate different values to different properties and thus result in different final scores from the evaluation process.

The PLANETS project builds on the Utility analysis work by integrating it into the PLATO Preservation Planning Tool, a web-accessible system for measuring and evaluating the performance of preservation activities against stated requirements and goals. (<http://www.ifs.tuwien.ac.at/dp/plato/>) The project has defined four main groups of characteristics: object, record, process and costs. In recognition that requirements vary across settings, it is recommended that as many stakeholders as possible are involved in the definition of requirements, from producers, curators and consumers to IT staff, domain experts, managers, and lawyers. The tool is still in development and will eventually integrate with registries and services for file format identification, characterisation and preservation actions.

Analysis of significant properties studies

In recent years it has become increasingly evident that a renewed study on the topic of significant properties was necessary, to gain a better understanding of the significant properties of various object types that institutions must maintain. To address this need the JISC funded the InSPECT project and four studies that would investigate the significant properties of several object types, including vector images, moving images, learning objects and software. These projects have been informed by the ‘Performance model’ and associated methodology created by the National Archives of Australia [14], as well as related work that has been performed previously.

Although the various significant properties studies share a common objective, they each developed specific

methodologies for the identification and interpretation of significant properties, partially based on archival diplomatics, utility analysis, records management and other discipline specific standards (e.g. the SPeLOs [1] project was informed by web-based e-learning practices and the Significant Properties of Vector Images study [7] was influenced by the Computer Graphics Reference Model).

One of several recommendations identified during the course of a workshop on the topic of significant properties was that the outputs of these projects should be mapped onto a common model to identify similarities and differences [11]. The final section of this paper will provide a comparison of the significant properties identified by the four recent JISC-funded studies. This work will enable the recognition of common themes between different objects based on their complexity (e.g. a software package and a raster image) and content type (still images, moving image). In addition, the outputs of each study may be merged to correct shortfalls in the coverage of each study. For example, the analysis of composite objects, such as Learning objects may be informed by analysis at a lower level, through use of the outputs of the studies into moving images or sound [12].

To begin to analyse the significant properties of the objects a conceptual framework is required. The study on the Significant Properties of Software [16] recognised the FRBR (Functional Requirements for Bibliographic Records) as being potentially useful for analysing different layers of a resource. FRBR is a conceptual entity-relationship model that represents the ‘products of intellectual or artistic endeavour’ at four layers of analysis: Work, Expression, Manifestation, and Item. In practical use, these layers may be equated to a Record, version of the Record, a variant of the version (e.g. an moving image object saved as an AVI and MPEG2; two variants of software compiled for Microsoft Windows and Linux); and Object that represents a single example of the work (e.g. a AVI file located on a user computer). However, to use the FRBR model as a basis for analysing significant properties, we must introduce a fifth entity, Component that represents one or more constituent parts of an object (e.g. an audio bit-stream in a moving image; a file in a software or learning object package).

1. Record

The Record is the top-level entity that equates to FRBR Work, The National Archives’ concept of a Record [23], or software ‘Package’. Several elements may be identified that indicate the significant properties for the Record entity in the studies on software [16], learning objects [1] and moving images [8] that describe the digital resources.

	<i>Software</i>	<i>Learning Objects</i>	<i>Moving Images</i>	<i>Vector Images</i>
<i>Context</i>	package name, keywords, purpose, Functional Requirement,	learning object classification, contextual, creator/Contributor, Description (Interactivity level, type, keywords) Educational Context, Metadata (catalogue type, references, subjects)	title	-
<i>Context: Rights</i>	provenance/owner	Rights management		

Table 1: significant properties for the Record/Work entity

The information specified for the Record entity is informed by an archival diplomatics and records management methodology. The metadata is useful for establishing the chain of custody and provenance of the digital resource and may assist with its location and retrieval in a digital archive. However, it is provided for the purpose of completeness and is not considered to be relevant for the purpose of preservation to maintain access to the digital resource, in part or whole.

2. Expression / Version

A FRBR Expression is a realization of the intellectual work in a specific form. This may equate to different versions of an object containing updated or changed content (e.g. a learning object that is used for teaching in 2008 and later modified for the same course in 2009) or functionality (e.g. a software package that provides a new user interface, import/export option, or other features). Matthews et al (2008) identifies 17 entities that may be recorded for each software version. In addition, descriptive information in the Learning Objects and Moving Images study may be identified that are relevant for each version of an object.

	<i>Software</i>	<i>Learning Objects</i>	<i>Moving Images</i>	<i>Vector Images</i>
<i>Context: descriptive</i>	Version identifier, Functional description, Input format, output formats, Description of the algorithm used, API description, Software specification	LO classification, Educational context, Validator record, Author record, Creation date, Title, Learning Assembly	Title	-
<i>Context:rights</i>	Licence	Digital		

		Rights management		
<i>Technical Environment</i>	Software dependencies, Architectural dependencies, Hardware dependencies			

Table 2: significant properties for the Expression entity

The properties that are attributed to the Expression share a common theme, indicating specific contextual information that describes the function for which it has been created (e.g. a learning object for use in learning and teaching; a software tool for creation and processing of data) and its use by a Designated Community. The list of associated items specified in the study of software is not considered to be significant properties. However, the existence of documentation is a key component in understanding a software tool and recompiling or re-implementing it for a different environment².

3. Manifestation

A FRBR Manifestation is the embodiment of an expression in a particular medium or format. For example, the encoding of a moving image resource in the Apple Quicktime format or as a series of TIFF images, or the compilation of software code for Microsoft Windows or Linux systems. It is likely that Representation Information will be created for each manifestation, to interpret and render the digital resource in an appropriate technical environment. In the context of significant properties, the studies of Software and Learning Objects have identified several properties that may be categorised with the Manifestation entity:

	<i>Software</i>	<i>Learning Objects</i>	<i>Moving Images</i>	<i>Vector Images</i>
<i>Context: description</i>	Variant notes	Learning Unit classification, Digital object datatypes, reusability		-
<i>Context: rights</i>	licence			
<i>Structure</i>	software dependencies; configuration (software)	Delivery		-
<i>Behaviour</i>		Look and feel Delivery		
<i>Technical Environment</i>	platform (software); operating system (software).	interoperability		

² It is less common for researchers to create similar documentation for other types of digital object. Digital archives, such as the UK Data Archive and the Arts & Humanities Data Archive recommend that resource creators document the digital outputs that they produce.

	Compiler (software); hardware dependencies (software);			
--	--	--	--	--

Table 3: Significant properties for the manifestation entity

The Manifestation properties describe the technical composition of the digital resource. At this level of analysis, there is the potential for confusion between Representation Information and Significant Properties. Notably, the classification of environment properties is a matter for discussion, particularly in relation to software packages. However, other elements are simpler to interpret as a significant property. The ‘Look and Feel’ and ‘Reusability’ elements incorporate aspect of the technical composition, but use them as the basis for specifying the allowed usage of the digital resource.

4. Item

An FRBR Item is a single instance of a manifestation. For example, a learning object or software package that is stored in a digital repository or on a user’s computer. It is equivalent to a software ‘Download’ or ‘installation’ [16]. A recipient may be provided with Representation Information to support its rendering and use or a description of significant properties to describe the content of the digital object. The majority of information provided with an item will have been created for each manifestation and, as a result will not require description at the item level. However, some object types may require the recording of information that indicate the digital rights and usage of the digital object in a specific environment (table 4).

	<i>Software</i>	<i>Learning Objects</i>	<i>Moving Images</i>	<i>Vector Images</i>
<i>Content</i>	-	-	No. of streams	-
<i>Context</i>	Licensee, Conditions, Licence code		Creation date	-
<i>Structure</i>	File relationships	Relationship between constituent parts (files, metadata)	Relationship between constituent parts (bitstreams)	-
<i>Technical Environment</i>	Environment variables, IP address, Hardware address			

Table 4: Significant properties for the Item entity

The Significant Properties of Software study has identified six properties that are distinct from those specified for the Expression or Manifestation entities. These indicate the licensee that is the user of the software; an individual licence tailored to the use of the particular item and user;

and hardware and software configurations that are distinct to the environment in which it will be used (e.g. the software can be used only if a specific IP or MAC address is defined). Similar requirements are not specified in the remaining three significant properties studies, though it is theoretically possible that a Learning Object, moving image, or vector image could be imprinted with a watermark or digital signature that is linked to a specific user. The location of the rights and environment properties is a matter for discussion. Although the study indicates that the properties are significant at the item-level, it may be better represented as a manifestation that has been tailored to the requirements of a specific user.

5. Component

A Component represents a unit of information that forms a logical group. The term is used by The National Archives [23], InSPECT [12] and Significant Properties of Software [16] projects to represent one or more sub-sections that, when aggregated and processed correctly will form the Item as a whole. It may be applied to several artefacts, including an audio bit-stream in a moving images file, a text paragraph in a HTML page and a shape in a vector graphics diagram. Significant properties that are defined for the component entity describe characteristics of the information content or the environment in which the content may be reproduced [13]. Each of the four studies identify information specific to the content type that they were responsible for analyzing:

	<i>Software</i>	<i>Learning Objects</i>	<i>Moving Images</i>	<i>Vector Images</i>
<i>Content</i>	-	text	duration	text
<i>Context</i>	functional description input format, Output format, Program language, Interface, Error handling			
<i>Structure</i>		-		
<i>Behaviour</i>		-		
<i>Rendering</i>	Algorithm	Text (format, character encoding, layout, fonts, colour) Animation (colours, frame rate, speed)	Gamut, Frame height, frame width, pixel aspect ratio, frame rate interlace	point, open path, closed path, , object, inline object, shape
<i>Behaviour</i>		-		
<i>Tech Environ-</i>	hardware depend-	-	compression ratio,	

<i>ment</i>	encies, library dependencies, package dependencies		codec	
-------------	--	--	-------	--

Table 5: Significant properties for the Component entity

The component entity is key for maintaining access to and use of the information object. The projects have recognized a range of technical properties that perform similar functions for each object type – recreation of the text, raster image and vector image of the object. However, it is questionable if elements classified under the Environment heading are properties of the information object or data object.

Conclusion

This paper has provided a definition of significant properties and outlined their role in a digital preservation strategy. It has highlighted criteria for their evaluation, through consideration of the requirements of those that have an investment in the availability and use of digital research, as well as work being performed in the international digital preservation community to assist institutions with the task of understanding and evaluating significant properties. The review of projects and institutions that have made some contribution to the development of digital preservation strategies suggests that there is a great interest in the identification, analysis and extraction of significant properties. However, the distinct methodologies adopted by each JISC project suggest that further work is necessary to encourage adoption of the Utility Analysis and Digital Diplomatics methodologies. The mapping of the significant properties to the FRBR entity-relationship model proved to be a useful exercise for understanding the disparate approaches taken by project and has highlighted similarities and differences between the properties for each object type. On the basis of the results obtained, it is evident that there remains some difference in the understanding of properties that may be categorized as significant for the information object and those that may be classified as Representation Information and that further work is necessary to map the significant properties of an information object onto a conceptual and practical model in a consistent manner.

We have yet to reach the stage where a researcher or academic in an institution is able to define the significant properties of their digital research without ambiguity. It is expected that ongoing work being performed by InSPECT, PLANETS and CASPAR and other projects will provide a common methodology and tools for understanding significant properties. In particular, work should be performed that maps the significant properties of an information object onto a conceptual and practical model in a consistent manner.

References

- [1] Ashley, K. Davis, R & Pinsent, E. 2008. Significant Properties of e-Learning Objects (SPeLOs), v1.0. http://www.jisc.ac.uk/whatwedo/programmes/programme_preservation/2008sigprops.aspx
- [2] Brown, A. 2008. White Paper: Representation Information Registries. http://www.planets-project.eu/docs/reports/Planets_PC3-D7_RepInformationRegistries.pdf
- [3] Bearman, D. & Trant, J. 1998. Authenticity of Digital Resources: Towards a Statement of Requirements in the Research Process, *D-Lib Magazine*, June 1998. <http://www.dlib.org/dlib/june98/06bearman.html>
- [4] CASPAR Project. n.d. Cultural, Artistic and Scientific Knowledge for Preservation, Access and Retrieval. <http://www.casparpreserves.eu/>
- [5] Cedars Project. 2002. Cedars Guide To : Digital Collection Management. <http://www.leeds.ac.uk/cedars/guideto/collmanagement/>
- [6] Chen, Su-Shing. 2001. The Paradox of Digital Preservation. *Computer*, vol. 34, no. 3, pp. 24-28, Mar., 2001. <http://doi.ieeecomputersociety.org/10.1109/2.910890>
- [7] Coyne, M et al. 2007. The Significant Properties of Vector Images. http://www.jisc.ac.uk/whatwedo/programmes/programme_preservation/2008sigprops.aspx
- [8] Coyne, M. & Stapleton, M. 2008. The Significant Properties of Moving Images. http://www.jisc.ac.uk/whatwedo/programmes/programme_preservation/2008sigprops.aspx
- [9] Digital Preservation Testbed. 2003. From digital volatility to digital permanence: Preserving text documents. <http://www.digitaleduurzaamheid.nl/index.cfm?paginakeuze=185>
- [10] Grindley, N. June 2008. The Significant Properties of Digital Objects. http://www.jisc.ac.uk/whatwedo/programmes/programme_preservation/2008sigprops.aspx
- [11] Hockx-Yu, H & Knight, G. 2008. What to Preserve?: Significant Properties of Digital Objects. *International Journal of Digital Curation*, Vol 3, No 1 (2008) <http://www.ijdc.net/ijdc/article/view/70>
- [12] Knight, G. 2008a. Framework for the Definition of Significant Properties. <http://www.significantproperties.org.uk/outputs.html>
- [13] Knight, G. 2008b. Significant Properties Data Dictionary. <http://www.significantproperties.org.uk/outputs.html>
- [14] Heslop, H. Davis, S. Wilson, A. 2002. An Approach to the Preservation of Digital Records. http://www.naa.gov.au/Images/An-approach-Green-Paper_tcm2-888.pdf
- [15] MacNeil, H. et al. Authenticity Task Force Report. http://www.interpres.org/book/interpres_book_d_part1.pdf
- [16] Matthews, B. et al. 2008. The Significant Properties of Software: A Study, Version 5.7. http://www.jisc.ac.uk/whatwedo/programmes/programme_preservation/2008sigprops.aspx
- [17] PLANETS Project. n.d. Preservation and Long-term Access through NETworked Services. <http://www.planets-project.eu/>
- [18] Potter, M, 'Researching Long Term Digital Preservation Approaches in the Digital Preservation Testbed', RLG DigiNews (June 2002) <http://digitalarchive.oclc.org/da/ViewObjectMain.jsp?fileid=0000070519:000006287741&reqid=3550>
- [19] Rauch, C. Strodl, S. & Rauber, A. 2005. Deliverable 6.4.1: A Framework for Documenting the Behaviour and Functionality of Digital Objects and Preservation Strategies. http://www.dpc.delos.info/private/output/DELOS_WP6_d641_final_vienna.pdf
- [20] RLG. 2002. Trusted Digital Repositories: Attributes and Responsibilities. <http://www.oclc.org/programs/ourwork/past/trustedrep/repositories.pdf>
- [21] Rothenberg, J. & Bikson, 1999. T. Carrying Authentic, Understandable and Usable Digital Records Through Time: Report To the Dutch National Archives And Ministry of the Interior. http://www.digitaleduurzaamheid.nl/bibliotheek/docs/final-report_4.pdf
- [22] The British Academy & The Publishers Association. 2008. Joint Guidelines on Copyright and Academic Research Guidelines for researchers and publishers in the Humanities and Social Sciences. <http://www.britac.ac.uk/reports/copyright-guidelines/index.html>
- [23] The National archives, n.d. http://www.nationalarchives.gov.uk/electronicrecords/seamless_flow/default.htm
- [24] Wilson, A. 2007. Significant Properties Report. http://www.significantproperties.org.uk/document/s/wp22_significant_properties.pdf

Towards a Curation and Preservation Architecture for CAD Engineering Models

Alexander Ball and Manjula Patel
UKOLN, University of Bath, Bath. BA2 7AY.

Lian Ding

IdMRC, Dept. of Mechanical Engineering, University of Bath, Bath. BA2 7AY.

Abstract

For many decades, computer-aided design (CAD) packages have played an important part in the design of product models within the engineering domain. Within the last ten years, however, the increasing complexity of CAD models and their tighter integration into the workflow of engineering enterprises has led to their becoming the definitive expression of a design. At the same time, a paradigm shift has been emerging whereby manufacturers and construction companies enter into contracts to take responsibility for the whole lifecycle of their products – in effect, to sell their product as a service rather than as an artefact. This makes necessary not only the preservation of the product's design, but also its continuing intelligibility, adaptability and reusability throughout the product's lifecycle. The CAD models themselves, though, are typically in closed formats tied to a particular version of an expensive proprietary application prone to rapid obsolescence. While product lifecycle management (PLM) systems deal with some of the issues arising from this, at present it is not possible to implement a comprehensive curation and preservation architecture for CAD models, let alone the other forms of engineering information.

In order to fill in some of the gaps in a possible architecture, we have developed two tools to aid in the curation and preservation of CAD models. The first is a preservation planning tool for CAD models: a Registry/Repository of Representation Information for Engineering (RRoRIE). The tool uses Representation Information, as defined by the Open Archival Information System (OAIS) Reference Model, to advise on suitable strategies for migrating CAD models to archival or exchange formats. The second – Lightweight Models with Multilayered Annotations (LiMMA) – is an architecture for layering non-geometric information on top of a geometric model, regardless of the format used for the geometric model. We envision this architecture being used not only to create flexible, lightweight archival representations of model data, but also to facilitate better information flows between a design team and the rest of the extended enterprise.

Introduction

Within the engineering industry, Computer Aided Design (CAD) has grown steadily in importance since its introduction in the mid-1950s (Bozdoc 2004). Originally used to aid in the production of design drawings, CAD can now define a design more clearly than two dimensional drawings ever could, and within the past decade has started taking over as the definitive expression of a design. With the corresponding rise of Computer Aided Manufacturing (CAM) and Computer Aided Engineering (CAE) systems, not to mention Enterprise Resource Planning (ERP), Customer Relationship Management (CRM) and Supply Chain Management (SCM) systems,

the potential for CAD models to be integrated with processes across a product's lifecycle is just starting to be realized.

The primary purpose of a CAD model is to represent the physical geometry of a design, typically in three dimensions. There are two different methods by which conventional CAD models represent the geometry of a product. Constructive Solid Geometry (CSG) constructs models as a combination of simple solid primitives, such as cuboids, cylinders, spheres, cones, etc. Boundary representations (B-rep), in contrast, represent shapes by defining their external boundaries: structured collections of faces, edges and vertices (McMahon and Browne 1996). Compared to CSG, B-rep is more flexible and has a much richer operation set, and so has been widely adopted in current commercial CAD systems. One of the ways in which B-rep models can be made highly expressive is through use of freeform surface modelling. This is where complex surface curvatures are represented using mathematical functions – such as Non-Uniform Rational B-Spline (NURBS) or Bezier surfaces – or approximations thereof.

CAD models can express more than just geometry, though. Most common CAD systems, whether using CSG and B-rep representations, can represent parts in terms of 'features', which encapsulate the engineering significance of the part as well as its geometry. Such features are often defined parametrically, allowing variations on the same basic part to be used throughout the model with little repetition of design data. Features are used not only for product design and definition, but also for reasoning about the product in a variety of applications such as manufacturing planning (Shah and Mäntylä 1995). While features are useful when coming to interpret a design, many are provided by vendors and/or embedded within CAD systems, making it hard to exchange the non-geometric information between systems. Additionally, features tend to be written from the designer's point of view, and may not fit the viewpoints of engineers in other parts of the extended enterprise.

The integration of CAD systems with other computerized systems in the manufacturing and in-service engineering phases is a significant part of Product Lifecycle Management (PLM), which aims to allow organizations to manage their products from conceptualization to disposal in the most efficient way possible. PLM is becoming increasingly important as organizations enter into more through-life contracts with their customers. Indeed, the extent to which customers are preferring to use a service model for acquiring products, particularly from the aerospace, defence and construction industries, has led some authors to describe this in terms of a paradigm shift (Davies, Brady, and Tang 2003;

Oliva and Kallenberg 2003). The product-service paradigm places a number of requirements on CAD, not least that the product data be kept intelligible, adaptable and reusable throughout the product's lifecycle. When considering the lifespan of some of the products – of the order of thirty or more years – this is not an insignificant challenge, especially given the rate of change of CAD software.

The CAD software industry is intensely competitive, with market forces driving rapid functional and performance improvements. While this has obvious benefits, it also has negative consequences. The ways in which the improvements are implemented cause conflicts not only with implementations on other CAD packages – and indeed with other types of systems – but also with those of earlier versions of the same CAD package. With little interoperability or backwards compatibility, and rapid turnover of software versions, CAD models can become unreadable within the timespan of three to ten years. That is not to say that CAD translation tools do not exist – they do – but due to the nature of the task they are not altogether reliable: in 2001 the manual correction of translated CAD data cost the aerospace, automotive, shipbuilding and tooling industries an estimated US\$74.9m in the US alone (Gallaher, O'Connor, and Phelps 2002).

Even leaving aside the preservation issues, there are barriers to using CAD in a PLM context. Every participant in the collaborative enterprise throughout the whole product lifecycle is expected to share product information – the staff in various departments within a company, partners, contractors/subcontractors, service providers and even customers – and CAD models carry most of the important information and knowledge. On the one hand, the cost of CAD packages makes it infeasible for staff outside the design team(s) to have access to the models. On the other hand, companies are naturally unwilling to share full product models that include commercially sensitive information, especially with temporary partners, with whom collaborative protocols are not established and who may at other times be competitors.

Furthermore, current CAD models are 'resource-heavy', and restrict information transmission between geographically distributed applications and users. The file size of a relatively simple component (e.g. a crankshaft) could be over 1MiB in one leading CAD system. Hundreds of such components may be included in a product such as a car, leading to very large storage requirements for models and restricting the options for their communication.

In the remainder of this paper, we report the state of practice with regard to PLM systems. We then present our proposed additions to PLM architecture to better cater for the curation and preservation of product model data. Finally, we present in more detail the set of significant properties of product model data used by our proof-of-concept tools and give our conclusions.

Product Lifecycle Management

Engineering organizations of reasonable size are likely to use a PLM system for managing their data. PLM systems offer a number of different functions, for instance: file storage (typically with version control, access permission control, simple on-access format conversions), cross-file linkages (e.g. bills of materials generated directly from CAD models), cross-system linkages (typically with ERP, CRM, and SCM systems), portals for various activities across the lifecycle (e.g. simulation analysis, maintenance log manage-

ment) and facilities for collaboration, both within lifecycle stages and between them. A number of PLM systems use lightweight formats – simple 3D formats that miss out much of the richness of a full CAD format – for communicating design information across the enterprise, and many claim to enforce compliance with various regulatory and certificatory requirements (Registration, Evaluation and Authorization of Chemicals; Six Sigma Quality; etc.).

While current PLM systems are certainly highly functional software environments, and do contain features pertinent to curation, they do not have any particular emphasis on preservation. None of the major PLM offerings (Dassault Systèmes, Siemens, SofTech, etc.) have integrated tools for preservation planning, monitoring when data storage media need to be refreshed, monitoring file format obsolescence, and so on. With some functions, such as wholesale migration from one CAD system and format to another, this is because the operation would be so complex, extensive and infrequent that it would need to be handled by a specialist team using specialist tools. With others, such as choosing appropriate lightweight formats for particular applications, it is because the PLM system architecture is only designed to support one option. Thus in order to fully support the curation and preservation of engineering documentation, additional tools are needed.

Proposed architecture

General framework

Within the digital library and digital preservation communities, several curation and preservation environments have already been developed.

PANIC (Preservation Web services Architecture for New media and Interactive Collections) is a semi-automated preservation environment developed by the University of Queensland (Hunter and Choudhury 2006). Its aim is to support three particular aspects of preservation: capture and management of preservation metadata, monitoring for format obsolescence, and interacting with preservation Web services. The architecture is modular, with separate local services for capturing and storing metadata, checking for obsolescence, discovering Web services, selecting Web services and invoking Web services. It relies on separate, probably external, registries for file formats and preservation Web services.

CRiB (Conversion and Recommendation of digital oBject formats) is a similar environment developed by the University of Minho (Ferreira, Baptista, and Ramalho 2007). It includes local services for detecting the formats of ingested materials, checking for format obsolescence, determining suitable alternative formats for ingested materials, determining suitable migration pathways, recording details of available preservation services, invoking preservation services, and evaluating the success or otherwise of preservation actions to inform future decisions.

PLANETS (Preservation and Long-term Access through NETworked Services) is a European Union funded project looking at practical preservation strategies and tools (Farquhar and Hockx-Yu 2007). One of its deliverables is a modular preservation environment; among other things, the environment consists of: Plato, a preservation planning tool; a testbed for evaluating preservation approaches; a software emulation environment; a tool for designing automated preservation workflows; a set of modules for carrying out

automated preservation workflows; a file format characterization registry; a preservation action registry; and a registry of preservation services.

It is clear that all three examples have much in common in terms of their architecture and the services they provide, and that these services are largely if not entirely absent from current PLM systems. That is not to imply that *all* of these services would be especially useful in the engineering context. For example, an obsolescence notifier would likely be of limited use as for large quantities of data within the organization, obsolescence comes about solely as a result of planned software upgrades rather than through environmental changes. Similarly, the migration of CAD models between major CAD formats is not a process to automate lightly, although other types of engineering documentation – reports, spreadsheets – may benefit from this sort of approach.

Another aspect that we feel deserves greater examination is the way PLM systems handle the communication of CAD data across the extended enterprise. Lightweight formats have particular advantages over full CAD formats, in that they are typically fairly simple, well documented and free from restrictive licences; this in turn means that it is relatively inexpensive to write software to support them, which means that such software is usually offered at little or no cost, and can be run across a number of platforms. All these things combined mean that they will likely remain readable for considerably longer than full CAD models. These advantages have not escaped CAD vendors, especially those who also produce PLM software, and a number have created their own. Because of this, there is a trend for PLM systems to support just one lightweight format for design review processes and the like. For example, Siemens Teamcenter uses JT, while Dassault Systèmes’ PLM offerings use 3D XML. This is unfortunate, as different lightweight formats have different characteristics that make them particularly suited to specific use cases. Furthermore, feeding back information from later in the lifecycle is typically achieved through an entirely different set of functions, meaning that the benefits of tying, for example, in-service maintenance records directly to the original CAD models – in order to inform future design choices – are left unexploited in current PLM implementations.

The architecture we propose would add to PLM systems the following functions: a registry of format characteristics, a registry of format migration services, a registry of (evaluations of) preservation actions, and a preservation planning tool based on top of these three registries. We also propose that PLM systems should adopt a more flexible, modular and consistent approach to communicating design information throughout the extended enterprise, the better to aid the curation of engineering information.

To this end we have developed two proof-of-concept systems, demonstrating how some of these functions may be implemented. The first, LiMMA (Lightweight Models with Multilayered Annotations), is a system for representing CAD models using lightweight geometric models supplemented with layers of XML-encoded information. The second, RRoRiE (Registry/Repository of Representation Information for Engineering), is a simple preservation planning tool that incorporates a registry of format characteristics and a registry of migration software.

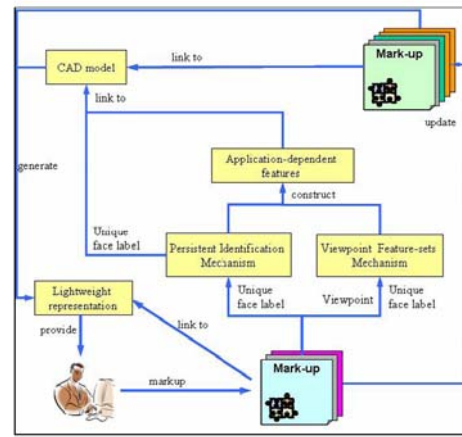


Figure 1: LiMMA -A Framework for the Annotation of CAD Models

LiMMA

LiMMA is not a single application or platform, but a series of individual tools based around a common XML schema and workflow. The premise behind it is that the same geometric model can exist in a number of different formats: full CAD formats, lightweight visualizations or exchange standards like STEP (ISO/TS 10303-203:2005) or IGES (US Product Data Association 1996). If extra information is added to the model in any one of these formats, and if that information is to be used to the widest possible extent, it ought to be visible in every other format, but this is problematic for at least three reasons: a) this would involve regenerating each version of the model every time information is added, b) different formats treat non-geometric information in different ways, and c) it would probably involve designing custom format translators. The solution in LiMMA is not to change the models at all but to store the information as annotations in a separate XML file and layer those annotations on top of the model using a system of persistent references (see Figure 1). Thus LiMMA consists of a series of plugins and viewers that allow one to interact with the annotation files whilst viewing the model, the system of persistent references used by the plugins and viewers, and the workflow of moving models and annotation files around the extended enterprise.

The multilayering of annotations in LiMMA is a way of offering additional flexibility and increasing the efficiency of the system. Not all the annotations will be of interest to everyone in the organization, and some may be confidential to a small group of engineers. By storing annotations in several different files according to access permission and interest groups, one can ensure that everyone receives all and only the annotations that they are allowed to see and that are of interest to them. The segregation of annotations into different files does not affect their usability as they are all layered on top of the model at once.

LiMMA has the potential to improve information flows throughout the product lifecycle. At the *design* stage designers can embed and share design rationale, meanwhile geographically distant design teams may collaborate on the same design using lightweight formats that preserve the exact geometry of the model. Any additional design information not recorded by the lightweight format – such as materials

and finishes – could be communicated using the annotation files. Similarly, an annotated lightweight/exchange version of the full model could be submitted to regulatory bodies for inspection, without either party having to invest in full CAD translations or multiple CAD package licences. Finally, the organization’s customers could be provided with lightweight models (using approximate geometry in order to protect the organization’s intellectual property) and function-related annotations. Similar models could be used as marketing materials to attract further customers.

By the *production* stage, the design has been finalized and lodged in the PLM system. A copy in a lightweight or exchange format, with accompanying annotations providing the additional design information and semantics to enable later re-editing, should also be kept in case the original model cannot reliably be opened when it is next needed. The CAD package in use by the design team and the Numerical Control (NC) software in use by the production engineers do not need to be so tightly integrated if the NC programmes can be generated from lightweight formats with exact geometry and manufacturing-related annotations. These could also be used by production engineers to feed back comments to the designers.

Once the product has reached its *in-service* phase, lightweight models with approximate geometry and annotations relating to disassembly and reassembly could be supplied to maintenance engineers, enabling them to have access to the design while inspecting the product. Inspection results could be marked up directly onto the model, allowing these results to be fed back to the designers as annotations. In this way, when the model is next opened for redesign or upgrade, any systematic in-service issues with the existing design can be spotted immediately and dealt with.

Finally, when the product has reached *end of life*, engineers could use a lightweight model with annotations relating to materials to determine which parts need to be disposed of in a controlled manner, and which can be recycled in some way; this type of information is also useful for input into future design and development.

So far, LiMMA plugins have been written in C/C++ and NX Open for the CAD package NX and in JavaScript for the 3D PDF viewers Adobe Acrobat and Adobe Reader, while a standalone LiMMA X3D viewer has been written using Java. The annotations are currently linked to the models by means of unique identifiers attached to surfaces within the model, but a parallel system of reference using co-ordinate sets is also under development.

RRoRIE

RRoRIE is primarily a planning tool, enabling information managers to explore the options available for converting CAD models into other formats, whether for contemporaneous exchange or for long term archiving. It does this by means of stored information about the capabilities of various formats and processing software with respect to certain significant properties (see Figure 2); the precise details are given in the section on significant properties below.

As well as simply allowing one to browse through the information contained in the self-contained repository, RRoRIE allows one to perform three different types of search on it. The first allows one to search for all the (known) formats that fit a chosen set of criteria with respect to significant properties. For each property, one can specify that it should be fully

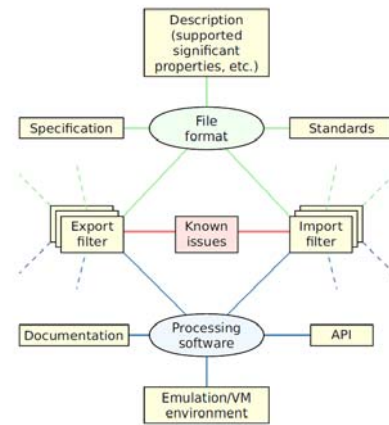


Figure 2: Capabilities of formats and conversion tools

supported, fully or partially supported, or not supported at all; otherwise it is not considered in the search. The second type of search calculates the possible migration paths between two formats, in a given number of steps or fewer. The third type of search allows one to specify a starting format and, as in the first type of search, a set of criteria for the final destination format. RRoRIE will then calculate a set of suitable migration pathways with the specified number of steps or fewer, and perform some simple ranking on them.

We anticipate RRoRIE being of use in at least the following scenarios: a) determining which lightweight formats would be suitable for which purposes when planning organizational LiMMA workflows and archival strategies; b) determining which tools or services to use to generate those lightweight formats from the full CAD models; and c) providing additional decision support when procuring a new CAD system to replace the existing one.

Further Work

There remain some outstanding issues with both LiMMA and RRoRIE that need to be resolved in order to fully demonstrate their usefulness. One of the use cases for LiMMA is for annotated lightweight or exchange formats to be used as an archival backup in case the original CAD model ceases to be readable. In order to prove this concept, we plan to determine if a set of annotations can be generated automatically from the non-geometric information in a full CAD model, and to assess the feasibility of reconstructing a full CAD model from an annotated lightweight model.

With RRoRIE, we intend to demonstrate how the Representation Information it stores may be synchronized with generic registries such as the Registry/Repository of Representation Information (RRoRI) developed by the UK’s Digital Curation Centre and the European CASPAR Project (Giaretta 2007). There is also plenty of scope for expanding RRoRIE to take account of more than just significant properties: openness of formats, price, availability and customizability of software, as well as evaluations of previous preservation actions.

Other areas of the proposed architecture we have not yet explored include the systematic evaluation of preservation actions, and the human and organizational issues associated with keeping the various registries up to date. Having argued

against the need for an obsolescence notifier in the given context, it may yet be useful to have a tool that measures the potential impact, with respect to the readability of files, of a proposed software upgrade or system change.

Significant Properties

The utility of the LiMMA system is predicated on the understanding that different viewpoints and different stages of the product lifecycle have varying uses and requirements for CAD models; some features of a model may be vital for one engineer and irrelevant for another. In other words, the *significant properties* of a CAD model vary between viewpoints and lifecycle stages. Significant properties are those aspects of a digital object which must be preserved over time in order for it to remain accessible, usable and meaningful (Wilson 2007, 8).

In the general case, what may be considered significant about an object depends partly on the nature of an object – for a Mercator projection map, true bearings are significant while areas are not, whereas for a sinusoidal projection map, areas are significant but bearings are not – and partly on the purposes to which it is put – such as whether one is concerned about a graph’s underlying data or its aesthetics. The latter dependency means that conceivably *any* property of an object may be significant to someone, so those entrusted with the preservation of the object have to prioritize the possible future uses of the object, and thereby the significant properties to preserve. In practice, for CAD models there are a limited number of ‘business’ uses (as opposed to academic uses) to which they could be put at present, although of course one cannot predict the future with any certainty.

For the purposes of constructing RRoRIfE, whose purpose is to compare different methods of expression and the processes of translating between them, we had to take a view on significant properties that was one step removed from the definition given above. We considered significant properties to be those aspects of a digital object which any new expression of that object must exhibit in order to fulfil its intended function while being faithful to the original; the notion of faithfulness is intended to encapsulate the given definition’s notion of preservation over time with respect to access, utility and meaning.

In the previous section on the proposed architecture, we outlined a number of use cases for CAD models. From these, several types of requirements can be identified:

- Some use cases require exact geometry, others approximate geometry.
- Some use cases require the modelling history;
- Some use cases require geometry-related metadata (tolerances, finishes, etc.);
- Some use cases require transmission of the model over a the Internet;
- LiMMA relies on persistent identification of (subsets of) geometry.

In the following subsections we present our working list of significant properties for CAD models, based on these requirements. The properties are structured in a hierarchy in order to take advantage of logical dependencies between them; for example, if a format is capable of analytically expressing an ellipse, it can certainly express a circle analytically. This allows for greater brevity when recording the expressiveness of different formats.

Geometry

There are two factors to consider when judging whether geometry expressed in one format may be expressed exactly in another format. The first is whether the entities used in the first expression have an equivalent in the second format, and the second is whether the conversion from the original entity to its equivalent can be done programmatically. The first of these can be determined relatively easily by comparing the basic entities supported by each format. Thus the first set of significant properties concerns geometric entities (Table 1).

These entities were compiled with reference to a previous study of the significant properties of vector graphics, and a number of different format specifications and software manuals (Coyne et al. 2007; ISO/TS 10303-203:2005 ; ISO/IEC 19775:2004 ; Shene 2007; Shene 1997; US Product Data Association 1996).

Geometric construction techniques

In order to build geometric entities into full CAD models, one or more construction techniques have to be used. The methods of construction available within a file format have a significant impact on its expressiveness, thus the second set of significant properties relates to these (Table 2).

One of the main distinguishing features of a format is whether it only allows parts to be made up of Boolean operations on solid objects (Constructive Solid Geometry), or whether individual surfaces can be used as well or instead (Boundary representation). There are further distinctions in the use of parametrically defined parts and construction history modelling. Finally, some formats have facilities for including several different versions of the same part; commonly this is used to speed up rendering – so viewers can render small or distant parts using low-fidelity meshes – but may be used to provide alternative organizational viewpoints on the same data.

Geometry-related metadata

The third set of significant properties is concerned with information about particular parts of the geometry, apart from shape information (Table 3).

In addition to the actual geometry, manufacturing and quality control processes require at the least geometric dimensioning and tolerancing information (giving the size of the various components and acceptable limits for errors), as well as information on the materials from which to make the components and the required finishes. Certain re-editing applications also require the preservation of the semantics associated with model ‘features’ (predefined geometry with established engineering meaning).

If a format provides a way of adding arbitrary metadata to a node in the assembly (a subassembly, part or perhaps surface), this can provide a way for additional geometry-related information to be embedded within the model. Even if the currently available software is unable to make use of this information, additional tools or plugins may be developed to interpret it.

Compression and identification

One of the factors that determine whether a format is likely to be suitable for transmission over the Internet, which may be necessary with geographically dispersed design teams, is whether a format tends to produce smaller file sizes. It was not considered within the scope of this project to devise a

Entity	Special case of
Point	–
Polyline	–
Line	Polyline
Conic arc	–
Elliptical arc	Conic arc
Circular Arc	Elliptical arc
Open composite curve	–
Closed composite curve	–
Ellipse	–
Circle	Ellipse
Polygon	–
Triangle	Polygon
Rectangle	Polygon
Square	Rectangle
NURBS curve (open or closed)	–
Rational Bézier curve	NURBS curve
Non-rational Bézier curve	Rational Bézier curve
Cubic Bézier curve	Non-rational Bézier curve
Quadratic Bézier curve	Cubic Bézier curve
Point cloud	–
Helix	–
Plane	–
Ellipsoid	–
Sphere	Ellipsoid
Cylinder	–
Cone	–
Cuboid	–
Cube	Cuboid
Torus	–
Mesh of surface segments	–
Mesh of tessellating triangles	Mesh of surface segments
Lofted surface	–
Ruled surface	Lofted surface
Translation surface	–
Normal swept surface	–
Polylinear swept surface	Normal swept surface
Extrusion surface	Polylinear swept surface
Swung surface	Normal Swept surface
Rotation surface	Swung surface
NURBS surface	–
Rational Bézier surface	NURBS surface
Non-rational Bézier surface	Rational Bézier surface

Entity	Special case of
Constructive Solid Geometry	–
Boundary representation	–
Trimmed surfaces (surfaces trimmed by boundary curves/surfaces)	–
Parameterized re-use of instances	–
Simple re-use of instances	Parameterized re-use of instances
Construction history modelling	–
Multiple alternative representations	–
Levels of detail	Multiple alternative representations

Entity	Special case of
Feature semantics	–
Material metadata	–
Geometric dimensioning and tolerancing	–
Dimensions	Geometric dimensioning and tolerancing
Assembly node metadata	–
Assembly hierarchy	–

Entity	Special case of
Field-wise compression	–
Stream-wise compression	–
Whole-file compression	–
Streaming	–
Identification of subassemblies	–
Identification of parts	–
Identification of surfaces	–
Identification of edges	–
Identification of vertices	–

reliable and fair metric for determining this quantitatively, so in lieu of this, our significant properties include various ways in which file sizes may be reduced. One method was mentioned above – re-use of a single part several times within a model – and the remainder are given here (Table 4). Another factor to be considered is whether the format allows streaming: allowing the file to be opened before it has been entirely transferred.

Finally, there is the matter of identification of the parts of a model. We are particularly interested in this from the perspective of using LiMMA, but there are other technologies which would benefit from being able to refer to identifiers within models.

Implementation in RRoRiFE

RRoRiFE uses two different XML schemata to store Representation Information, one for file formats and one for conversion processes; each schemata is based on the above ontology of significant properties.

The first schema relates to file formats and describes whether or not the format supports a particular property. As well as ‘full’ support and ‘none’, an intermediate value of ‘partial’ support is allowed, to indicate that support is limited in some way; for example, NURBS surfaces may be allowed, but only with 256 or fewer control points. In cases of partial support, explanatory text must be provided.

The second XML schema relates to conversion processes, grouped by software product. For each format conversion – and each optional variation of that conversion – the software is able to perform, the schema allows one to record how well the conversion preserves each property. Four levels of preservation are allowed. ‘None’ indicates that the property has never knowingly survived the conversion intact (most frequently because the destination format does not support the property). ‘Good’ indicates that the conversion has so far preserved examples of the property sufficiently well that it would be possible to reconstruct the original expression of the property from the new expression. ‘Poor’ is used when



Figure 3: User Interface to RRoRiFE

tests have found it at least as likely for the property to be corrupted or lost as it is to survive. Lastly, ‘fair’ is used in all other cases, alongside an explanatory note.

Where preservation is less than ‘good’, it is possible to record whether the property survives in a degraded form, and if so, whether this degradation always happens in a fixed way, a configurable way or an unpredictable way. For example, when moving from a format that supports NURBS to one that only supports tessellating triangles, there may be a fixed algorithm for approximating surfaces, or one may be able to specify how detailed the approximation is.

The hierarchy of the ontology has been programmed into RRoRiFE, so that it knows that if a format supports NURBS surfaces, for example, it also supports non-rational Bézier surfaces. It does not make these inferences, though, if the Representation Information file in question already contains a statement about the ‘child’ property. Figure 3 shows the GUI to RRoRiFE.

Conclusions

In this paper we have argued that CAD packages and PLM systems do not currently provide the functionality required for the preservation of engineering materials, nor for taking full advantage of potential information flows within organizations responsible for the full lifecycles of their products. We therefore propose the addition of several new components to the PLM system architecture: a system of lightweight models and layers of annotation, to facilitate easier and more far-reaching information flows; a registry of file format characteristics, to help determine the suitability of the formats for specific purposes; a registry of format migration software and services, and a registry of (evaluations of) preservation actions, to aid in planning migration strategies; and a preservation planning tool based on top of these three registries. In order to test the feasibility of these architectural additions, we have developed two proof-of-concept systems. LiMMA demonstrates how annotations stored in dedicated XML files may be layered on top of CAD models in a variety of formats, using application plugins and custom viewers via a persistent reference mechanism. These annotations may be passed around an organization independently of each other and used with any translation of the referent model. In addition, as they are simpler and better documented than full CAD formats,

lightweight formats are better suited to long term preservation, and some of the information lost in translation may be preserved instead as annotations. RRoRiFE demonstrates how information about the support that file formats and processing software have for the significant properties of CAD models can be used to support preservation planning decisions. There are, of course, still a number of issues to resolve with the proposed architecture, not least the human and organizational aspects of maintaining such a system, but we believe that it is promising, worth studying and developing further.

Acknowledgements

This work is supported by the UK Engineering and Physical Sciences Research Council (EPSRC) and the Economic and Social Research Council (ESRC) under Grant Numbers EP/C534220/1 and RES-331-27-0006. The Digital Curation Centre is funded by the UK Joint Information Systems Committee (JISC).

References

- Bozdoc, M. 2004. The history of CAD. Available from: <http://mbinfo.mbdesign.net/CAD-History.htm> (2008-02-01).
- Coyne, M.; Duce, D.; Hopgood, B.; Mallen, G.; and Stapleton, M. 2007. The significant properties of vector images. Technical report, Joint Information Systems Committee.
- Davies, A.; Brady, T.; and Tang, P. 2003. *Delivering Integrated Solutions*. Brighton: SPRU/CENTRIM.
- Farquhar, A., and Hockx-Yu, H. 2007. Planets: Integrated services for digital preservation. *International Journal of Digital Curation* 2(2):88–99.
- Ferreira, M.; Baptista, A. A.; and Ramalho, J. C. 2007. An intelligent decision support system for digital preservation. *International Journal on Digital Libraries* 6(4):295–304.
- Gallagher, M. P.; O’Connor, A. C.; and Phelps, T. 2002. Economic impact assessment of the international Standard for the Exchange of Product model data (STEP) in transportation equipment industries. RTI Project 07007.016, National Institute of Standards & Technology, Gaithersburg, MD.
- Giaretta, D. 2007. The CASPAR Approach to Digital Preservation. *International Journal of Digital Curation* 2(1):112–121.
- Hunter, J., and Choudhury, S. 2006. PANIC: an integrated approach to the preservation of composite digital objects using Semantic Web services. *International Journal on Digital Libraries* 6(2):174–183.
- ISO/IEC 19775:2004. Information technology – Computer graphics and image processing – Extensible 3D (X3D).
- ISO/TS 10303-203:2005. Industrial automation systems and integration – Product data representation and exchange – Part 203: Application protocol: Configuration controlled 3D design of mechanical parts and assemblies (modular version).
- McMahon, C., and Browne, J. 1996. *CADCAM: From Principles to Practice*. Wokingham: Addison-Wesley.
- Oliva, R., and Kallenberg, R. 2003. Managing the transition from products to services. *International Journal of Service Industry Management* 14(2):160–172.
- Shah, J. J., and Mäntylä, M. 1995. *Parametric and Feature-Based CAD/CAM*. Chichester: Wiley.

Shene, C.-K. 1997. A user guide to the surface subsystem of DesignMentor. Available from: <http://www.cs.mtu.edu/~shene/COURSES/cs3621/LAB/surface/surface.html> (2008-08-07).

Shene, C.-K. 2007. Introduction to computing with geometry notes. Available from: <http://www.cs.mtu.edu/~shene/COURSES/cs3621/NOTES/> (2008-08-07).

US Product Data Association. 1996. Initial Graphics Exchange Specification (IGES) v5.3. Withdrawn standard ANS US PRO/IPO-100-1996, ANSI.

Wilson, A. 2007. InSPECT significant properties report. Technical report, Arts and Humanities Data Service/National Archives.

Evaluating Strategies for the Preservation of Console Video Games

Mark Guttenbrunner, Christoph Becker, Andreas Rauber, Carmen Kehrberg

{guttenbrunner,becker,rauber,kehrberg}@ifs.tuwien.ac.at

Vienna University of Technology, Vienna, Austria

<http://www.ifs.tuwien.ac.at/dp>

Abstract

The amount of content from digital origin permanently increases. The short lifespan of digital media makes it necessary to develop strategies to preserve its content for future use. Not only electronic documents, pictures and movies have to be preserved, also interactive content like digital art or video games have to be kept “alive” for future generations. In this paper we discuss strategies for the digital preservation of console video games. We look into challenges like proprietary hardware and unavailable documentation as well as the big variety of media and non-standard controllers. Then a case study on console video game preservation is shown utilizing the Planets preservation planning approach for evaluating preservation strategies in a documented decision-making process. While previous case studies concentrated on migration, we compared emulation and migration using a requirements tree. Experiments were carried out to compare different emulators as well as other approaches first for a single console video game system, then for different console systems of the same era and finally for systems of all eras. Comparison and discussion of results show that, while emulation works in principle very well for early console video games, various problems exist for the general use as a digital preservation alternative. It also shows that the Planets preservation planning workflow can be used for both emulation and migration in the same planning process and that the selection of suitable sample records is crucial.

Introduction

Video games are part of our cultural heritage. The public interest in early video games is high, as exhibitions, regular magazines on the topic and newspaper articles show. Games considered to be classic are rereleased for new generations of gaming hardware as well. However with the rapid development of new computer systems the way games look and are played changes rapidly. As original systems cease to work because of hardware and media failures, methods to preserve obsolete video games for future generations have to be developed. When trying to preserve console video games, one has to face the challenges of classified development documentation, legal aspects and extracting the contents from original media like cartridges with special hardware. Special controllers and non-digital items are used to extend the gaming experience. This makes it difficult to preserve the look and feel of console video games.

The term “video game” can refer to different kinds of electronic games where a person (“player”) plays a game primarily produced by a computer and usually presented

on some kind of display unit. These games are played on systems which have not been designed primarily for gaming (e.g. personal computers, mobile phones, digital cameras, classic home computers) as well as on systems made specifically for gaming (e.g. consoles connected to a TV, hand held consoles, arcade machines).

The challenge to preserve diverse types of video games varies in many aspects such as used media for software, kinds of presentation, levels of known system architecture. This work concentrates on the preservation of console games. These are devices that are specially made for playing games where the system’s output is displayed on a television screen. Example console systems are Atari 2600, Nintendo Entertainment System (NES) and Sony PlayStation.

First this paper gives an overview of related work. Then we present the challenges and discuss the different strategies for digital preservation for console video games. Next we show a case study for the long-term preservation of console video games using different digital preservation strategies. Similar preservation planning case studies concentrated on migration. We compare emulation and migration using the Planets¹ preservation planning approach to evaluate alternatives using an objective tree. Finally we present the conclusions that can be drawn from the experiments.

Related work

In the last years migration and emulation have been the main strategies used in digital preservation. Lorie differs between the archiving of data and the archiving of program behavior. While the first can be done without emulation, it cannot be avoided for the latter (Lorie 2001). While this rigorous statement may be challenged if re-compiling or porting code to a different platform are viewed as a form of migration, emulation definitely plays an important role for the preservation of program behavior.

¹ Work presented in this paper is partially supported by European Community under the Information Society Technologies (IST) Programme of the 6th FP for RTD -Project IST-033789. The authors are solely responsible for the content of this paper. It does not represent the opinion of the European Community, and the European Community is not responsible for any use that might be made of data appearing therein.

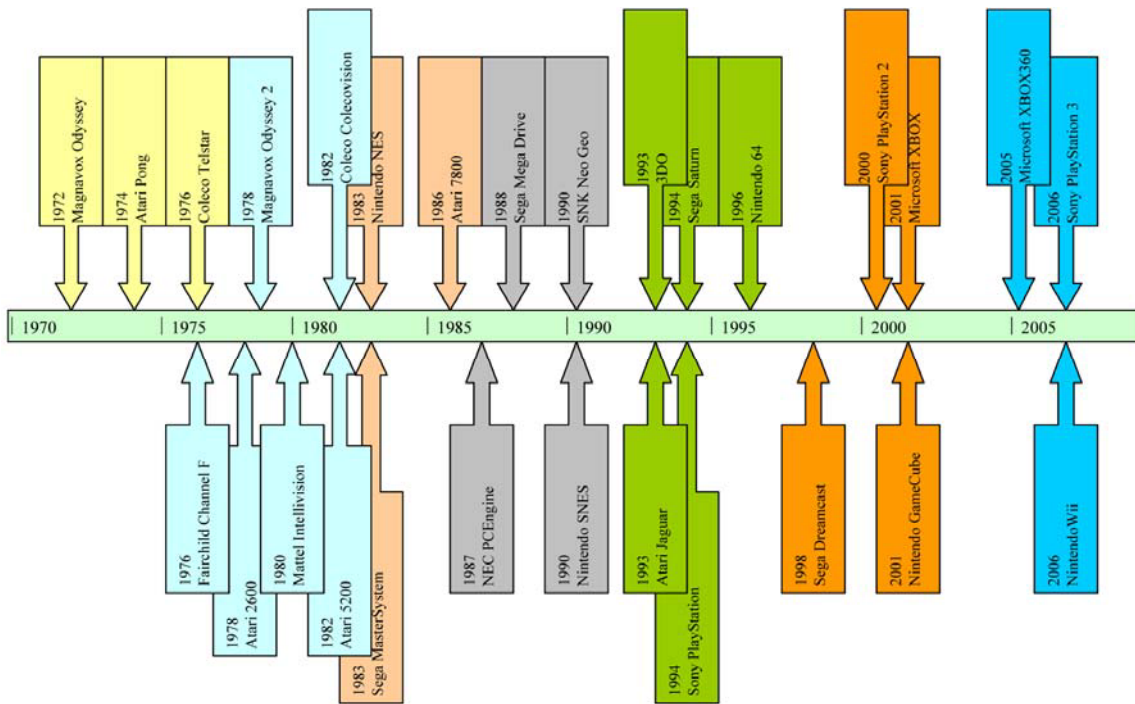


Figure 1: Timeline of release years for console video game systems. Systems of the same era are shown in the same color.

The concept of using emulation for digital preservation is to keep the data in its original, unaltered form and keep using the software originally used to display the data. This software has to be run on the operating system and the operating system on the hardware it was developed for. To keep this chain alive, an emulator for the original hardware is produced. Emulation can take place on different levels (software, operating system or hardware) as described in (Rothenberg 2000).

Several methods to establish emulation as a long term strategy for digital preservation are discussed in (Slats 2003). The concept of an Emulation Virtual Machine (EVM) was used for development of the Universal Virtual Computer (UVC) by IBM (van der Hoeven, van der Diessen, and van en Meer 2005). An approach to developing an emulator on a hardware level is discussed as a conceptual model in (van der Hoeven and van Wijngaarden 2005) as modular emulation. An emulator which uses the modular emulation approach (van der Hoeven, Lohman, and Verdegem 2007) is under development in the Planets project. Planets is a project developing services and technology to address core challenges in digital preservation co-funded by the European Union under the Sixth Framework Programme (Farquhar 2007).

A practical experiment on how to use emulation to recreate interactive art is presented in (Jones 2004).

The Planets preservation planning approach used for this case study is described in detail in (Strodl et al. 2007). Becker et. al. present case studies on sample objects of interactive multimedia art from the collection of the Ars Electronica² in (Becker et al. 2007).

² <http://www.aec.at>

Challenges

When preserving video games, one is faced with two different tasks: preserving the video game system and preserving the games themselves. The requirements and challenges for digitally preserving console video games are partially very different to those of preserving static documents and even video games on other systems like personal computers, home computers and arcade machines.

This case study concentrated on strategies for systems which had substantial market shares and are considered as major systems. Figure 1 shows a time-line of release years. Most of the results of this work are applicable to other console systems as well.

Numerous specific challenges have to be faced when preserving console video games. Unlike personal computers or early home computers, the exact specifications of console video game systems and development documentation for game developers are usually confidential.

Console video games have always been offered on various types of media which in most cases cannot easily be read on standard computer hardware. The most common media include ROM-cartridges which potentially also contained extra hardware besides a microchip storing the data. Optical media and on-line content are mainly used for the last generations of console systems.

While in many digital preservation appliances the user experience plays a minor role, it is considered the central aspect with interactive fiction like video games. To enhance the gaming experience especially for early video games, screen or controller overlays were used. These

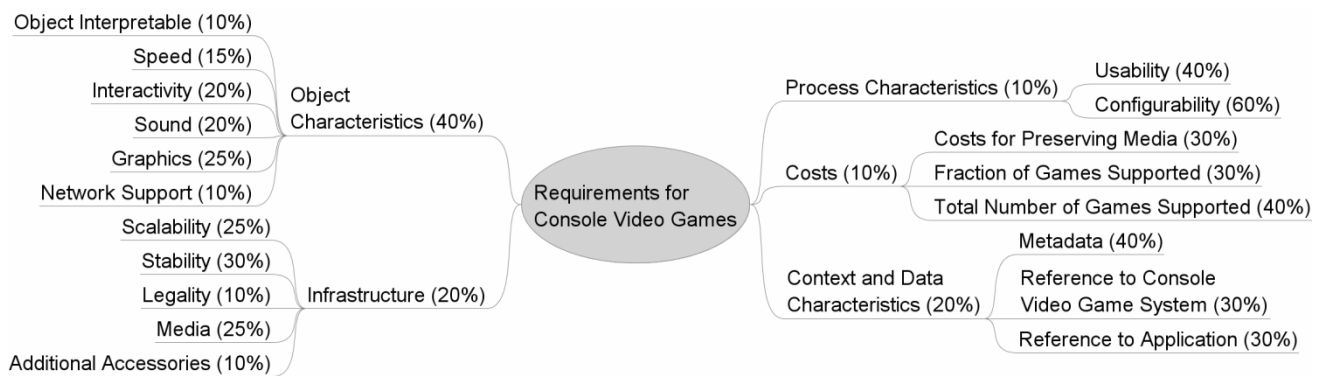


Figure 2: Requirements tree for console video games with importance factors (first two levels only).

overlays were applied to either the screen to enhance the visual impression of the image or to the controller to explain button layouts. The experience with some games lies in the use of a specially designed controller. Therefore it is necessary to find a way to recreate the game-play experiences with these games as close to the original as possible.

To preserve video games in any other way than keeping the original hardware and media, legal issues have to be addressed. It is necessary to constitute the responsibility for the preservation of digital data. Legal deposit laws should be extended to include digital data. The legal situation would have to be adjusted to carrying out the actions needed for digital preservation.

Strategies

Several strategies for preserving digital content are listed in the UNESCO Guidelines for the Preservation of the Digital Heritage (Webb 2005). Applied to the preservation of console video games, they can be summarized as follows.

The *Museum Approach* is not a long term preservation strategy as console video game systems are usually built from custom manufactured parts which cannot be replaced once broken.

Only screen shots (or non-digital videos) of video games could be preserved with the *Print-to-Paper Approach*, this does in no way preserve the dynamic look and feel of interactive content. This will for most applications not be a sufficient preservation strategy for video games.

Backwards Compatibility, the strategy to let consumers use games of earlier systems on newer generation models has been a successful commercial strategy since the third generation of video games (e.g. adapter to use Atari 2600 games on a Colecovision console (Herman 2001)). However once a manufacturer goes out of business, the games are no longer supported by a future system. As soon as the media is defective the contained video game is lost for preservation, too.

Code re-compilation for new platforms is one approach to *Migration* also known as *Source Ports*. Unavailable source code, the proprietary hardware of console video game systems and the usually very platform dependent code make it next to impossible to migrate a game to a new platform.

Another migration strategy is the approach to create a video of the game. Although all interactivity is lost, this gives a good representation of the original visual and audible characteristics of a game and can even be used as a future reference for recreating the game in an interactive way.

Simulation is another strategy that can be used for the preservation of console video games. Reprogramming a game might be possible for very early and simple games without knowing the original code. For more complex games and systems with more than just very few games this is either not possible or too costly compared to other alternatives.

For console video games *Emulation* may be the most promising solution, as most systems have to be well documented for video game software developers to write games. Only one piece of software (the emulator) has to be written to run the library of all games for a console system instead of having to deal with every piece of software for a given system.

Evaluation of Strategies

We evaluate various different solutions for preserving console video games. For this we used the Planets Preservation planning workflow for making informed and accountable decisions on a preservation strategy. The planning tool PLATO which supports this workflow as well the detailed results of this case study are available online via the PLATO homepage³.

The setting that was used for the case study is a future library. It is expected to have a mandate similar to a national library to collect published digital games and make them available for the public over a long term. The major goals are an authentic look and feel of the preserved games, easy accessibility, stable solutions for a long term preservation, and high compatibility with all games for the systems.

To achieve wide representation of significant properties to be preserved, we chose three games for each of the video game console systems we wanted to preserve as sample records. We selected sample records with these considerations in mind:

³ <http://www.ifs.tuwien.ac.at/dp/plato/>

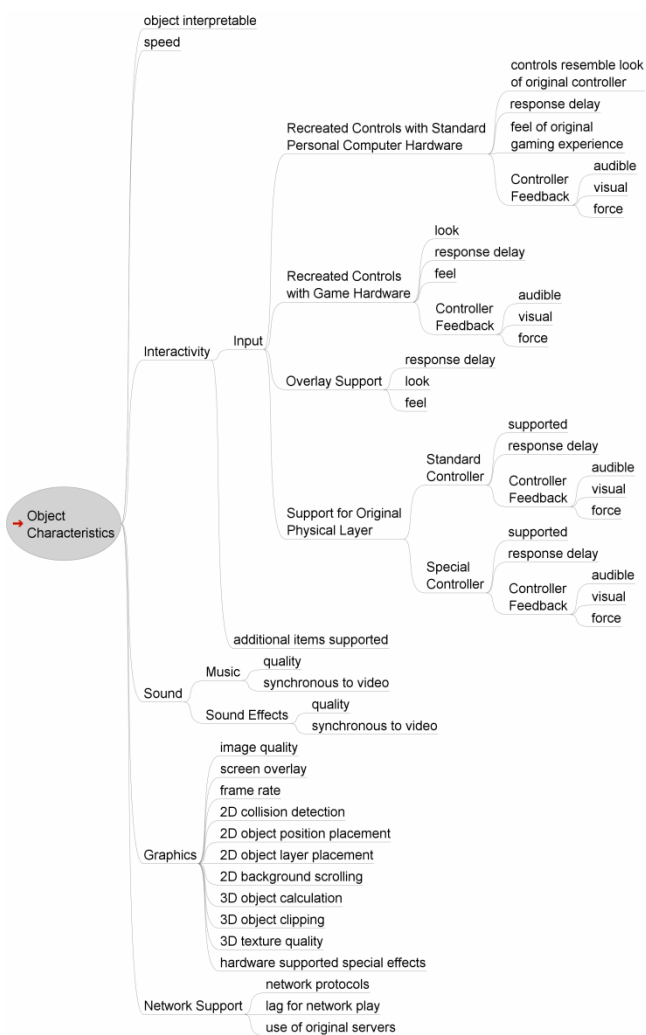


Figure 3: Object characteristics in the requirements tree.

- one major game for the system which most likely attracts the public's highest attention
- one game that uses special controllers to evaluate the feel aspect
- one of the games that make most intensive use of hardware-specific functions

The sample records chosen for the evaluated systems are shown in Table 1.

The requirements were collected and structured into an objective tree along the following five main categories (Figure 2):

Object Characteristics - The significant properties of video games are reflected in the visual, audible and interactive characteristics of the reproduced object. They are shown in the sub-tree in Figure 3. Visual aspects are divided into overall image impression as well as 2D and 3D features of the evaluated sample games. Sound aspects are divided into music and sound effects. Speed and the support of additional aspects like network support were tested as well. The typical scale that was used for measuring the degree to which an object requirement was met is:

- feature not applicable for this sample record
- feature not supported by the alternative
- feature supported but severe errors noticeable
- feature supported, errors noticeable but not affecting game-play
- no errors noticeable

The interactive requirements are used to measure look, feel and feedback not only for the use of standard PC components supported, but also for the support of special controllers and possible support for the original controls. Additional game items like overlays or off-screen game pieces have been considered in the requirements tree as well.

Process Characteristics - Part of this branch of the requirements tree is the configurability of a solution. It represents how easy it is to set configurations for a specific game and the system itself. Usability is the second sub-branch. It shows how straightforward and quickly games can be selected and if context specific data can be displayed with the game.

Infrastructure - This branch, depicted in Figure 4, gives the ability to measure information about how scalable and stable a solution is. The values in this part of the objective tree are used to collect data about the long term suitability of a solution. Details about the kind of development of a solution, the type of media supported and legal implications as well as expected support for a solution are considered.

Context and Data Characteristics - This branch describes the support of metadata of the game and necessary configuration options either with the solution or bundled with the game data.

Costs - This includes costs involved in retrieving data from the original media as well as costs for the preservation solution itself per supported game.

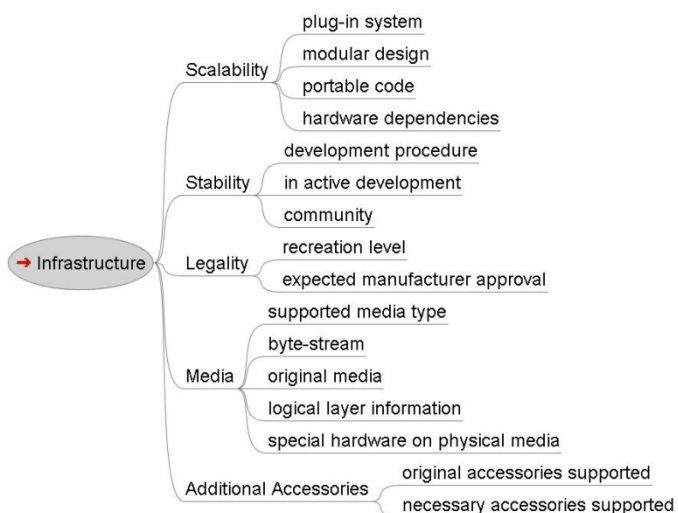


Figure 4: Infrastructure characteristics in the requirements tree.

In total, the tree contains 81 leaf criteria. We set importance factors to weight these leafs (Figure 2). On the top level the highest value was assigned to the object characteristics. Infrastructure for a long-term sustainability as well as the support of metadata was also assumed to be of a high importance. Costs were considered as important as well. Process characteristics are of less importance, as it is not necessary to browse very quickly through lots of games.

Three experiments were defined: Different alternatives for one system (Super Nintendo Entertainment System, also known as Nintendo SNES) were compared to check for differences in the performance of representatives of the same strategy as well as differences between strategies. Different alternatives for systems from the same generation (Nintendo SNES, Sega Genesis, NEC PCEngine, SNK Neo Geo) were evaluated to find out if some systems are better supported than others. Finally different alternatives for systems from all generations (Coleco Telstar, Philips G7000, Sega MasterSystem, Nintendo SNES, Atari Jaguar, Sony PlayStation 2) were evaluated to compare alternatives as systems evolved, i.e. whether a single emulator can show favorable performance across a range of systems.

Depending on the systems selected for evaluation suitable alternatives were chosen (Table 1). We selected emulation and simulation (where available) strategies as alternatives as well as a migration to video for comparison. Backwards compatibility and the museum approach were ruled out because they are short term approaches only. Source ports were not considered as source code is in general not available for games on the evaluated platforms.

The experiments were developed for a defined hardware and software setting, ran and evaluated. For every leaf in the tree the measured values were recorded for the three defined experiments with the selected systems and sample records.

Evaluation Results

This section presents the results of the evaluation procedure. We start with discussing the strategies used for the three experiments and an analysis of the aggregated results. We point out strengths and weaknesses observed and compare the different approaches that were evaluated.

Analyzing the evaluation results showed that for the two dedicated emulators chosen as alternatives for Nintendo SNES the results were very similar. Both were able to produce the visual and audible characteristics very well, even for games with additional hardware on the game cartridges. No metadata for the games are supported, and the emulators are written in platform-independent code for speed reasons. The multi-system emulator tested was not able to start one game with additional hardware at all and serious flaws on the produced images were visible for other sample objects (Figure 5). Portability is high due to platform independent code. Cost characteristics are good for the multi-system emulator, as a lot more games (for other systems) are supported. With the migration video-taping approach it was possible to

System	Alternatives	Sample Records
Nintendo SNES	ZSNES 1.51 SNES9X 1.51 MESS 0.119	Super Mario World Super Scope 6 Starfox
Nintendo SNES	video/audio grabbing with Hauppauge WinTV PVR and viewed with VLC 0.8.6c	Super Mario World
NEC PCEngine	MagicEngine 1.0.0. MESS 0.119	Bonks Revenge Gates of Thunder
Sega Mega Drive	Gens32 1.76 Kega Fusion 3.51	Sonic the Hedgehog 2 Daxside
SNK Neo Geo	NeoCD 0.3.1 Nebula 2.25b	Metal Slug Crossed Swords 2
Coleco Telstar	Pong 6.0 PEmu	Tennis
Philips G7000	O2EM 1.18 MESS 0.119	K.C. Munchin Quest for the Rings
Sega MasterSystem	Dega 1.12 Kega Fusion 3.51	Alex Kidd in Miracle World Space Harrier 3D
Atari Jaguar	Project Tempest 0.95 MESS 0.119	Doom Highlander
Sony PS2	PCSX2 0.9.2	Gran Turismo 3 Eye Toy Play

Table 1: Alternatives and sample records chosen for the experiments. All listed alternatives are emulation approaches except one migration video-taping approach for Nintendo SNES and simulation approaches for Coleco Telstar.

reproduce the look and sound perfectly, however the interactive element was lost. Metadata was supported by the file-format and the viewer that was used was open-source and platform-independent. Cost characteristics are very good for the video-taping approach as well, as it can be used for all games for all systems.

Similar results were observed for other systems of the same generation. Emulators were able to produce visuals and audible characteristics of the games well with bad infrastructure characteristics. For the multi-system emulators infrastructure characteristics were good, but not all games were playable.

The following results were observed for systems of different generations: The two simulators for Coleco Telstar were playable, but the feeling of the original paddle controllers was lost as only keyboard and mouse are supported. Infrastructure characteristics are bad, as none of the simulators is open source and development on both has been stopped. Support for Philips Videopac game pieces was not available in any of the emulators. The differences in infrastructure and costs between the multi-system and dedicated emulators are the same as observed before. Only one of the evaluated Sega MasterSystem emulators supported the 3D effect of one of the sample games. Atari Jaguar emulation was only partially working. The only available emulator able to play commercial games for the Sony PlayStation 2 was not able to produce in-game graphics for any of the sample objects. It did however support network functions



Figure 5: Screenshots of Super Mario World for the Nintendo SNES. Both pictures show the same scene. On the left is an image produced by ZSNES 1.51, on the right the same image as shown by MESS 0.119.

and provided the ability to use Sony's on-line service. Metadata was not supported in any case. Cost characteristics were better for later systems as more games were supported due to more available games per system.

According to the Planets preservation planning workflow the measured values were then transformed to a uniform scale of 0 to 5 with 0 being a value unacceptable for the use of an alternative and 5 being the best possible result. Values not applicable for a sample record or system are transformed to 5 to reflect an unchanged behavior compared to the original system.

The transformed values were then accumulated following the Planets preservation planning approach by weighted sum and weighted multiplication. This yields a ranking of the evaluated alternatives, reflecting their specific strengths and weaknesses. Three different emulators as well as the migration video-taping approach for preserving games for the Nintendo SNES video game console have been evaluated. The aggregated results can be seen in Table 2. Weighted Sum and Weighted Multiplication results for the alternatives separated into the top level branches of the requirements tree are shown in Figure 6. Minimal differences exist between the

dedicated emulators. The multi-system emulator has better results in infrastructure, but lacks compatibility to certain games using special hardware on the cartridge. The video approach has very good characteristics in almost all categories, but has to be eliminated because of missing interactivity in the object characteristics. If lack of interactivity was not considered critical, this would have been the optimal solution. It can also be a suitable back-up strategy for quick access or to verify future emulators' visual and audible compliance.

For systems of the same generation as the Nintendo SNES the results were similar. Dedicated emulators were better with object characteristics whereas multi-system emulators had better results in infrastructure and costs.

Simulators for very early consoles (the Coleco Telstar) had different approaches. While one was trying to enhance the visuals, the other stayed true to the original. Dedicated and multi-system emulators for consoles prior to the Nintendo SNES were almost equally good in reproducing visual and audible characteristics with better results on infrastructure for multi-system emulators. The evaluated emulators for systems of the last three generations of video game consoles were either not able to play commercial games yet or had low compatibility.

Alternative	Sample record	WS Sample	Mult. Sample	Mult. Total	WS Total
ZSNES 1.51	Super Mario World	3,45	2,75	3,28	2,68
	Super Scope 6	3,30	2,70		
	Starfox	3,38	2,78		
SNES9X 1.51	Super Mario World	3,43	2,82	3,31	2,70
	Super Scope 6	3,28	2,68		
	Starfox	3,38	2,78		
MESS 0.119	Super Mario World	3,56	2,88	2,68	0,00
	Super Scope 6	3,47	2,79		
	Starfox	2,47	0,00		
VLC 0.8.6c/MP4	Super Mario World	4,65	0,00	4,65	0,00

Table 2: Aggregated experiment results for preserving games for the Nintendo SNES (WS = Weighted Sum, Mult.= Multiplication). The highest values for each sample record as well as the highest ranked alternative are printed in bold.

Conclusions

In this work we used the Planets preservation planning approach to evaluate alternatives for the digital preservation of console video games. The same requirements tree was used to evaluate emulation as well as migration strategies. The case study showed that the Planets preservation planning workflow can be used to evaluate different strategies in one preservation planning process.

Furthermore it showed that the selection of sample records is especially crucial for emulation strategies and the archival of program behavior. While some sample records were reproduced flawlessly by an alternative, other sample objects could not be rendered at all by the same alternative. The results of the planning process are thus very dependent on the sample records. When doing preservation planning and considering emulation as a strategy, sample objects should be chosen with this fact in mind.

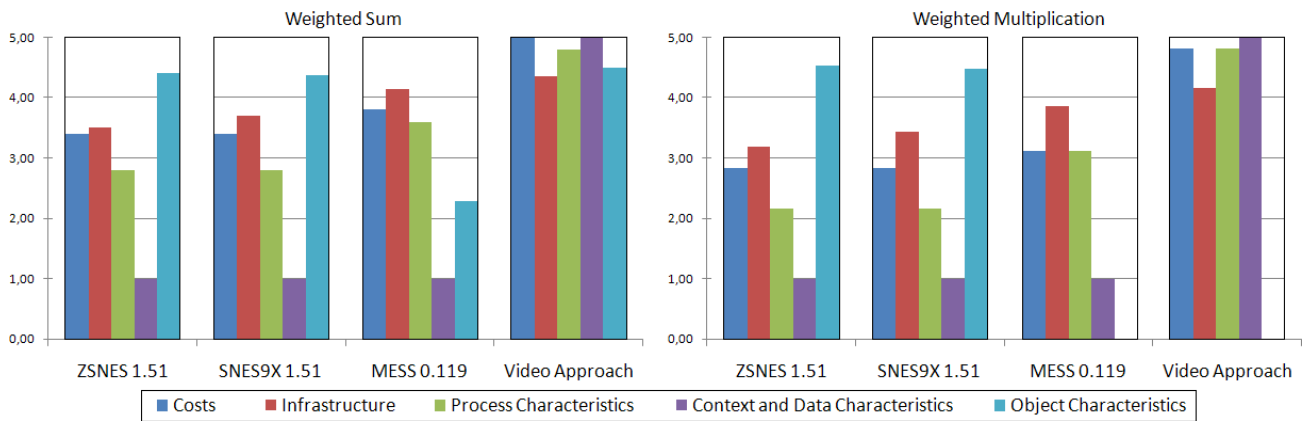


Figure 6: Aggregated results for the main categories in the requirements trees for Nintendo SNES preservation alternatives (weighted sum and weighted multiplication).

From the alternatives the migration to video approach showed very good results in most categories, but naturally completely failed in interactivity. The object characteristic results depended extremely on the chosen sample records with the evaluated emulators, however for the sample records with good object characteristics the interactive aspect of the games was still present.

The emulation alternatives had disadvantages in terms of infrastructure characteristics and metadata. Most emulators with high compatibility and good performance are not platform independent. Emulators supporting more than one system are usually modular and platform independent, but lack compatibility for certain games using special characteristics of the system. Metadata is not supported by the tested emulators, all expect raw binary streams of data. Compatibility and speed decreases dramatically for the emulation of modern systems. The feel aspect is only preserved well for games using standard controllers.

While all tested emulators were able to reproduce the original video or audio output to some extent, most are not usable without modification for digital preservation. Future work should focus on improving stability and metadata handling to provide a viable preservation solution for console video games. Work has also to be done in finding ways to recreating the original feel aspect of video games.

References

Becker, C.; Kolar, G.; Kueng, J.; and Rauber, A. 2007. Preserving interactive multimedia art: A case study in preservation planning. In *Asian Digital Libraries. Looking Back 10 Years and Forging New Frontiers. Proceedings of the Tenth Conference on Asian Digital Libraries (ICADL'07)*, volume 4822/2007 of *Lecture Notes in Computer Science*, 257–266. Hanoi, Vietnam: Springer Berlin / Heidelberg.

Farquhar, A., H.-Y. H. 2007. Planets: Integrated services for digital preservation. *International Journal of Digital Curation* 2(2).

Herman, L. 2001. *PHOENIX The Fall & Rise of Videogames - Third Edition*. Rolenta Press.

Jones, C. 2004. Seeing double: Emulation in theory and practice. The Erl King case study. In *Electronic Media Group, Annual Meeting of the American Institute for Conservation of Historic and Artistic Works*. Variable Media Network, Solomon R.Guggenheim Museum.

Lorie, R. 2001. A project on preservation of digital data. *RLG DigiNews* Vol. 5 (3). <http://www.rlg.org/preserv/diginews/diginews5-3.html#feature2>.

Rothenberg, J. 2000. *Using Emulation to Preserve Digital Documents*, Tech. Rep. Koninklijke Bibliotheek.

Slats, J. 2003. Emulation: Context and current status. Tech. Rep. http://www.digitaleduurzaamheid.nl/bibliotheek/docs/white_paper_emulatie_EN.pdf.

Strodl, S.; Becker, C.; Neumayer, R.; and Rauber, A. 2007. How to choose a digital preservation strategy: Evaluating a preservation planning procedure. In *Proceedings of the 7th ACM IEEE Joint Conference on Digital Libraries (JCDL'07)*, 29–38.

van der Hoeven, J., and van Wijngaarden, H. 2005. Modular emulation as a long-term preservation strategy for digital objects. In *5th International Web Archiving Workshop (IWA05)*.

van der Hoeven, J.; Lohman, B.; and Verdegem, R. 2007. Emulation for digital preservation in practice: The results. *International Journal of Digital Curation* Vol. 2 (2):123–132.

van der Hoeven, J.; van der Diessen, R.; and van en Meer, K. 2005. Development of a universal virtual computer (UVC) for long-term preservation of digital objects. *Journal of Information Science* Vol. 31 (3):196–208.

Webb, C. 2005. *Guidelines for the Preservation of the Digital Heritage*. Information Society Division United Nations Educational, Scientific and Cultural Organization (UNESCO) – National Library of Australia.

<http://www.unesco.nl/images/guidelines.pdf>.

Costing the Digital Preservation Lifecycle More Effectively

Paul Wheatley

The British Library
Boston Spa, Wetherby, West Yorkshire, LS23 7BQ, United Kingdom

paul.wheatley@bl.uk

Abstract

Having confidence in the permanence of a digital resource requires a deep understanding of the preservation activities that will need to be performed throughout its lifetime and an ability to plan and resource for those activities. The LIFE (Lifecycle Information For E-Literature) and LIFE² Projects have advanced understanding of the short and long-term costs in this complex area, facilitating better planning, comparison and evaluation of digital lifecycles.

The LIFE Project created a digital lifecycle model based on previous work undertaken on the lifecycles of paper-based materials. It applied the model to real-life collections, modelling their lifecycles and studying their constituent processes. The LIFE² Project has reviewed and refined the costing model and associated tools, making it easier for organizations to study, cost and compare their digital lifecycles in a useful way. New Case Studies provided useful practical experience of the application of these costing tools and brought the LIFE approach full circle by investigating the comparison of complex digital and analogue lifecycles. The Case Studies were able to elicit useful results, although digital preservation lifecycle costing remains a complex and involved process.

The LIFE Project

The LIFE Project was funded by JISC to explore the costing of digital preservation activities using a lifecycle approach. The project ran for 12 months, ending in April 2006. It was a collaboration between The British Library (BL) and University College London (UCL).

Background and Research Review

The Project began with a comprehensive review of existing lifecycle models and digital preservation costing activities (Watson 2005). The concept of lifecycle costing, which is used within many industries as a cost management or product development tool is concerned with all stages of a product's or process's lifecycle from inception to retirement. The review looked at applications of the lifecycle costing approach in several industries including construction and waste management, in order to identify, assess and potentially reuse an appropriate methodology.

It was within the Library sector that the greatest synergy and potential for adaptation to the digital problem area was found. A model for estimating the total cost of keeping a print item in a library throughout its lifecycle provided a useful starting point (Stephens 1988). Although developed for the paper world, there were interesting parallels between the stages of analogue and digital asset management that would subsequently prove useful. The original model was later extended to cover preservation costs (Shenton 2003). The lifecycle stages start with selection, acquisitions processing, cataloguing and press-marking and continue through to preservation, conservation, storage, retrieval and the de-accession of duplicates. Three key "life stages" were selected as useful reference points at which to calculate costs. Year 1 provided an indication of initial costs following the significant selection and acquisition stages. Year 10 represented a review point and possible technological change or surrogacy. Year 100 was chosen as the symbolic "long-term" point, useful for forecasting downstream costs. Building on the foundations of this primarily print-focused lifecycle approach, LIFE developed a costing model for digital materials.

The LIFE Model

The LIFE Model v1.0 (Ayriss, McLeod and Wheatley 2006) provided a content independent view of the digital lifecycle, breaking it down into Stages and Elements (see Figure 1). Each LIFE Stage represents a high-level process within a lifecycle that groups related lifecycle functions that typically occur or recur at the same point in time. These related functions are termed LIFE Elements. The LIFE model provided a common structure to which specific lifecycles could be mapped, enabling costing, analysis and comparison in a concise, readable and consistent manner.

The LIFE Methodology

LIFE implemented a simple methodology for the capture, calculation and recording of lifecycle costs. Key costs were identified for each element in the lifecycle. These might include equipment costs, setup costs and ongoing staff costs. An appropriate method of capturing these key costs was then identified and applied. Capital costs were averaged across their expected lifetime utilising the

Acquisition	Ingest	Metadata	Access	Storage	Preservation
Selection	Quality Assurance	Characterization	Reference Linking	Bit-stream Storage Costs	Technology Watch
IPR	Deposit	Descriptive	User Support		Preservation Tool Cost
Licensing	Holdings Update	Administrative	Access Mechanism		Preservation Metadata
Ordering & Invoicing					Preservation Action
Obtaining					Quality Assurance
Check-in					

Figure 1: the LIFE Model v1.0, showing the breakdown of Stages (across the top) and Elements (down the page)

number of objects that would be processed. Staff costs were captured using studies of the involved personnel and the time they spent on different tasks. Costs were simply projected over time based on present day value, without consideration for inflation. LIFE calculated costs for 1, 5, 10 and 20 years.

The LIFE Case Studies

Three case studies were chosen for the application and evaluation of the LIFE Model and Methodology. They were:

- Web Archiving at the British Library
- Voluntarily Deposited Electronic Publications (VDEP) at the British Library
- E-Journals at UCL

The resulting lifecycle costs and the full workings of how these costs were calculated can be found on the LIFE website (www.life.ac.uk).

The Generic Preservation Model

The Case Studies considered by the first phase of LIFE did not contain activities addressing the preservation of content, such as technology watch, preservation planning or migration. With no preservation processes to observe and cost, an alternative strategy had to be pursued. Attention was focused on the development of a model to estimate the long-term preservation costs. The work of

Oltmans and Kol (2005) provided a useful starting point on which to build a more detailed model. Desk research and various team review and evaluation work led to the creation of the Generic Preservation Model (GPM). The GPM takes as an input a basic collection profile and provides as output estimates of the costs of preserving that collection for a certain period of time.

The LIFE² Project

While the LIFE Project was felt to have made significant progress in this difficult problem area, the project team felt that there was still much to do in advancing our ability to accurately assess, cost and compare digital lifecycles. Although the first phase of the project had devised a useful approach and had provided some indicative costs and analysis in case study form, a more thorough test, review and strengthening of this approach was necessary.

The LIFE team successfully applied for funding for a second phase of the project (LIFE²), which began in March 2007 and ran for 18 months. The British Library and UCL again implemented the project but added a number of Associate Partners to develop new Case Studies.

Review and Further Application Of The LIFE Approach

The Project started by initiating an independent assessment of the economic validity of the LIFE approach to lifecycle costing which was undertaken by Professor Bo-Christer

Creation or Purchase	Acquisition	Ingest	Bit-stream Preservation	Content Preservation	Access
....	Selection	Quality Assurance	Repository Administration	Preservation Watch	Access Provision
....	Submission Agreement	Metadata	Storage Provision	Preservation Planning	Access Control
....	IPR & Licensing	Deposit	Refreshment	Preservation Action	User Support
....	Ordering & Invoicing	Holdings Update	Backup	Re-ingest	
....	Obtaining	Reference Linking	Inspection	Disposal	
....	Check-in				

Figure 2: the LIFE Model v2.0

Björk from Hanken, the Swedish School of Economics and Business Administration (Björk 2007). The report largely validated the approach taken by the LIFE team and provided a number of recommendations to help steer the second phase of the project in the right direction.

The LIFE Model and Methodology was then reviewed and updated by the project team, using the independent assessment, as well as feedback gathered from the wider digital preservation community, as a foundation for this work. This resulted in version 1.1 of the LIFE Model (Wheatley, et al. 2007).

The revised LIFE tools were applied to new LIFE Case Studies, two of which were conducted at Associate Partner sites:

- SHERPA DP, which examined the lifecycle costs of a preservation service
- SHERPA-LEAP, which studied lifecycle costs at the institutional repositories of Goldsmiths at the University of London, Royal Holloway at the University of London, and UCL (University College London)
- British Library Newspapers, which studied and compared both analogue and digital lifecycles at this National Library

A fourth Case Study that had planned to examine the costs of primary data curation was not completed due to staffing issues at the Associate Partner site.

Lessons learnt from the experiences of the Case Studies were fed back into the LIFE approach resulting in a further release of the LIFE Model as version 2.0. Full details of the Case Studies and their findings can be found in the LIFE² Project Final Report (Ayriss, et al. 2008) and key aspects of the LIFE approach that have enhanced our ability to cost digital lifecycles more effectively are discussed below.

LIFE² Developments

LIFE² invested considerable effort in developing the LIFE Model, Methodology and associated tools in order to improve the accuracy and consistency of the costing process, to simplify the work involved and to ensure that the results of lifecycle costing activities could be usefully applied.

An assessment of digital preservation costing objectives was undertaken, with the aim of identifying where the application of lifecycle costing data would be useful, and thus informing the development of the tools used to capture that data. Key objectives included:

- Identification of selective costs, such as repository running costs
- The cost of adding a new content stream lifecycle to an existing repository
- Evaluating the efficiency of an existing content stream lifecycle

- Assessing the impact of a new tool or a process change within an existing content stream lifecycle
- Comparison of similar lifecycles at different organisations
- Comparison of analogue and digital preservation

This assessment provided useful guidance in the development of the scope of lifecycle costing, which is addressed in more detail in the LIFE² Final Report.

The LIFE Model was revised following collation of a range of feedback on the LIFE¹ work. The LIFE team also liaised closely with the digital preservation costing team at the Danish Royal Library, State University Library and State Archives who provided invaluable comment and contribution as the Model was developed. The resulting release of the LIFE Model v2.0 provided a more detailed and more clearly defined picture of the digital lifecycle. Significant changes included clearer terminology, new lifecycle elements, particularly in Bit-stream Preservation, and more detailed definitions. As well as further description at the Stage and Element level, suggested Sub-element descriptions were included. These low-level lifecycle functions provide an indication of the scope and level of detail that would be useful to capture in a costing exercise, and most were found to be applicable for the lifecycles encountered in the Case Studies.

Conclusions

The experiences of implementing the Case Studies indicated that enhancements made to the LIFE Methodology, Model and associated tools have simplified the costing process. Mapping a specific lifecycle to the LIFE Model is not always a straightforward process. The revised and more detailed Model has reduced ambiguity. The Sub-element detail provides clearer guidance on the process of matching particular lifecycle processes to the LIFE Elements. The costing templates, which were refined throughout the process of developing the Case Studies, ensure clear articulation of both working and cost figures, and facilitate comparative analysis between different lifecycles. Despite these improvements, the addition of further detail to the Methodology would be desirable. This might include a tighter definition of the key processes and more guidance for users on the key costing procedures. While reviewing the LIFE Model, the team envisaged a categorization of cost types (e.g. capital, staff, development) and a more formal approach for capturing, costing and projecting them. Unfortunately, there was insufficient time to implement this. With the benefit of hindsight, it is clear that a more rigorous Methodology would have been useful, and should ideally have been prioritised over other developments.

Capturing the costs of lifecycles that are no longer actively ingesting digital objects proved to be problematic.

Although this was considered as a possible risk while planning the Case Studies in LIFE², it was not expected that this approach would be as time consuming as it turned out to be. Further difficulties were experienced in capturing a sufficient level of detail (with clear working) at the Associate Partner sites. As noted above, a more detailed methodology would have helped, but in contrast with the LIFE¹ Case Studies, it was clear that costing activities are far easier to lead within the managing organisation's own realm of responsibility. Far more effort was required to implement the LIFE² Case Studies than was expected, and this placed a considerable strain on project resources.

The complex nature of the lifecycles examined in the British Library Newspaper Case Study provided a thorough test of the LIFE approach for comparing and contrasting analogue and digital costs. The Case Study was able to elicit results that allowed some useful comparisons to be made, but the complexity involved highlighted that these analogue and digital mappings were very much in their infancy. The LIFE team is keen to further develop and explore our ability to compare and contrast analogue and digital lifecycle costs with the ultimate aim of informing the difficult digital versus analogue collection management decisions looming on the horizon.

Considerable progress has been made in costing the digital preservation lifecycle, despite the relatively small effort that has so far been directed at this complex and multi-faceted problem area. Since the start of LIFE¹, other new developments have emerged. A consortium of Danish organizations, including the National Library, State and University Library and the State Archives, are developing the LIFE Model for cross-institutional comparison of their digital preservation activities. JISC also funded a study into the costs of data curation, which utilized elements of the LIFE work (Beagrie, Chruszcz and Lavoie 2008). Despite these advances, digital preservation costing remains in its infancy and our current tools can provide us with indicative but not accurate digital preservation costs.

Moving forward our ability to cost the digital preservation lifecycle will require further investment in costing tools and costing models. Developments in estimative models will be needed to support planning activities, both at a collection management level and at a later preservation planning level once a collection has been ingested. In order to support these developments a greater volume of raw cost data will be required to inform and test new cost models. Organisations undertaking digital preservation activities are therefore encouraged to record costs as they proceed and where possible make their figures available to the wider community.

Looking ahead to LIFE³

A third phase of the LIFE work is currently under consideration. Initial proposals include a focus on development of an integrated toolset to both streamline the process of costing an existing digital lifecycle and estimate the cost of implementing a new lifecycle. The predictive tool would take as an input a simple profile of a new digital collection or content stream and a profile of the preserving organisation. The tool would then automatically process these profiles and estimate the costs for each lifecycle stage for a required timescale.

Acknowledgements

Thanks go to the LIFE Project Team and the many other staff of The British Library, UCL, the Centre for eResearch and the University of London, without whom the results of the Project would not have been realised.

References

Watson, J. 2005 *The LIFE project research review: mapping the landscape, riding a life cycle*. London, UK. <http://eprints.ucl.ac.uk/1856/1/review.pdf>

Stephens, A. 1988. *The application of life cycle costing in libraries*. *British Journal of Academic Librarianship* 3, 82-88

Shenton, H. 2003. *Life Cycle Collection Management* LIBER Quarterly 13, 254-272. <http://liber.library.uu.nl/publish/articles/000033/article.pdf>

McLeod, R., Wheatley, P., and Ayris, P., 2006. *Lifecycle information for e-literature: full report from the LIFE project*. LIFE Project, London, UK. <http://eprints.ucl.ac.uk/1854/>

Oltmans, E., Kol, N. 2005. *A Comparison Between Migration and Emulation in Terms of Costs*. *RLG Diginews* 9, <http://worldcat.org/arcviewer/1/OCC/2007/08/08/0000070519/viewer/file959.html#article0>

Björk, B.-C. 2007. *Economic evaluation of LIFE methodology*, LIFE Project, London, UK. <http://eprints.ucl.ac.uk/7684/>

Wheatley, P., Ayris, P., Davies, R., McLeod, R. and Shenton, H. 2007. *The LIFE Model v1.1*. LIFE Project, London, UK. <http://eprints.ucl.ac.uk/4831/>

Ayris, Davies, McLeod, Miao, Shenton, Wheatley, 2008. *The LIFE² Final Project Report*. LIFE2 Project, London, UK. <http://www.life.ac.uk/2/documentation.shtml>

Beagrie, N., Chruszcz, J., and Lavoie, B. 2008. *Keeping research data safe: a cost model and guidance for UK Universities*. JISC, London, UK. <http://www.ndk.cz/dokumenty/dlouhodobachrana/keeping-research-data-safe-a-cost-model-and-guidance-for-uk-universities>

Risk Assessment; using a risk based approach to prioritise handheld digital information

Rory McLeod

The British Library, Digital Preservation Team
St Pancras, 96 Euston Road, London NW1 2DB, UK
rory.mcleod@bl.uk

Abstract

The British Library (BL) Digital Library Programme (DLP) has a broad set of objectives to achieve over the next few years, from web-archiving to the ingest of e-journals through to mass digitisation of newspapers and books. These projects are decided by the DLP programme board and are managed by the wider corporate governance structure which includes our legal deposit responsibilities. As part of this work it was identified by the Digital Preservation Team (DPT) that a significant number of handheld media (CDROM, DVD, Tape) within the BL collections may be at increased risk of obsolescence or decay due to the increased time they may spend on handheld media. The DPT and DLP agreed that an assessment should be undertaken and the results used to help prioritize future ingest.

The DPT conducted this risk assessment exercise in order to assess the condition of the BL digital collections, identify strategies to mitigate those risks, and recommend and plan actions to be taken. A risk assessment methodology based on the AS/NZS 4360:2004 standard was applied in a representative manner across these collections.

The Risk Assessment concluded that the BL's digital collections face an array of risks that will require action on a number of fronts. Almost all of the hand held (physical carrier) collections were assessed to be at high risk.

The greatest and most imminent threat of loss is from media degradation. Failure rates for discs within the collections have reached high levels (up to 3%).

Additionally substantial quantities of digital objects are stored as single copies only, on handheld media in danger of decay. This stark warning was illustrated by many examples of disc decay that have been encountered and is backed up by the evidence from external research into handheld media lifetimes. Digital content will continue to be lost unless action is taken now. The report made a number of specific recommendations to mitigate the highest risks facing the BL's digital collections.

These include:

- Secure collections that are currently stored on handheld media as a matter of urgency. (Move the collections from CD/DVD etc)
- Perform further assessment to gain a better understanding of the media failure rates across the different collections
- Address the root causes of a number of the risks facing the collections, by streamlining and enhancing standards, check-in procedures and other policy issues

In order to achieve this a number of organisational changes have had to be undertaken that will eventually become measurable benefits.

Using a Risk Based Approach and its organisational impact

Overview

This paper will describe the organisational context within which the BL's 2007 risk assessment should be understood. It is not a technical overview of the methodology or the results as this information can already be found through the final report available at the BL's Digital Preservation website¹. What this paper will do is relate how the process of undertaking such a risk assessment informs the organisational and change management activity that is required to fundamentally shift the perception of digital preservation activity within an institution.

This paper will describe the different levels of organisational involvement required to undertake such risk based activity. The paper will also profile the awareness raising that has resulted in this piece of work becoming one of the most significant piece of analysis so far done by the BL's DPT, and how it has become a catalyst for various follow-up work scheduled for 2008/09. The aim of this paper would be to describe just how beneficial such a piece of work can become to running the business of preservation within a memory institution.

The start point

The BL's DPT is only three years young, incidentally our birthday coincides with the hosting of this years IPRES 2008 event so it is a good time for us to review past work and to think about what the next three years will hold for the vision of digital preservation at the BL and how the risk assessment forms a vital part of this work.

As part of our activities as the DPT we are determined to make sure that our work is representative of the Library's entire digital holdings. This means that as well as the broad corporate programmes (outlined in the abstract) we should focus at least some part of our effort on the material within the collections that does not have a

prioritised timeframe for ingest into our digital library system (DLS). This approach allows us to focus on our growing hand-held media collections and on digital content that may be outside the scope of current DLS work. By using the risk assessment we inform both our DLP and Collection Management strategies as well as providing a practical and measurable information source about the overall state of our collections.

In order to achieve this we first found that we had to undertake a detailed analysis of the BL collections and tie this analysis to the risk assessment following an approach based on international standards for risk management.

The 2007 risk assessment is based on the AS/NZS 4360:2004² Risk Management standard. This standard defines a seven-step approach to risk management:

Communicate and consult

Communicate and consult with internal and external stakeholders as appropriate at each stage of the risk management process and concerning the process as a whole.

Establish the context

This step sets the scene for the analysis. Stakeholders are identified, and the objectives of the stakeholders and the organization as a whole are established. If possible, measurement criteria are established so that the impact risk has on these objectives can be determined.

Identify the risks.

In this stage, the risks—that is, *what can go wrong*—are enumerated and described.

Analyze the risks

This step covers the evaluation of the impact of the risks, and the likelihood of those risks. The evaluation may be qualitative (an event may be “likely”, “unlikely”, “inevitable”, etc.) or quantitative (“a hard drive failure will occur on average once every 100,000 operational hours”), or some combination of the two.

Evaluate the analysis

At this stage, negligible risks might be discarded (to simplify analysis), and evaluations (especially qualitative evaluations) adjusted. The risks are compared to the objectives of the organization, allowing a ranked list of risks to be constructed.

Treat the risks

The options to address the risks are identified, the best option chosen, and implemented. This may include “taking no action” if no risk is sufficient.

Monitor and review

It is necessary to monitor the effectiveness of all steps of the risk management process. This is important for continuous improvement. Risks and the effectiveness of treatment measures need to be monitored to ensure changing circumstances do not alter priorities.

With the exception of steps 3 and 6 you can see how the methodology refers you to any organisational policies and strategies that may exist and asks you to make reference to them before you proceed. This is a good sanity check before starting as your organisation may not value this activity highly or this may in itself identify a gap in the strategic plan that is worth investigation.

The Risk assessment also provides a way of tying these strategic plan to the operational objectives of the business, for example the BL has a very clear digital strategy³ so for us it was very easy to balance the effort required for this work against the strategy of the organisation. This involved a small scoping study where we worked through points 1 and 2 of the methodology to establish the context. At this point we allowed ourselves a little time to develop the idea within our team using our own department plus our steering group as a mechanism to approve (in this case) our approach.

Communicate and Consult

The first stage of implementing a risk based assessment of digital content is to outline your communication plans and identify your key stakeholders. The BL is an organisation that has geographical challenges due to its multiple sites plus it has challenges of size. This is an organisation of some 2000 people and making your voice heard within such a business is a critical part of the success you can expect. As such the DPT outlined a clear communications strategy to assist, we initially took a top down approach and used our Executive steering group which involves our CEO and a number of Directors and Heads of Department, we presented our plan for the execution of the risk assessment and then allowed a period of time to address concerns raised by this group. The types of queries asked prior to our start were, who will undertake the work? What will be the time commitments in each department? And how much will this help us address or prioritise our digital content, why is this different to the 2003 study?

Our answer were

The DPT will be the primary resource and allocations for time have been given to the two key people involved. Each Department Head should support this and allocate us some time from one member of their team.

When we examined the 2003 risk assessment we concluded:

- Having the object isn't enough
- Knowing the format of the object or its content isn't enough
- You need software to use it, a computer to run it on
- The functionality and access of the object can intimately depend on the details of the environment, most of which we don't have.
- The organisation and business needs to change to support any attempt at e-collection management.

This information was presented to our own team in order to achieve good understanding of the unknowns that we were trying to address.

Internal communications mechanisms such as the Intranet and our staff publications were used to explain why the DPT were undertaking this study and what the benefits would be.

Additionally a questionnaire was compiled as part of the communication plan to be sent to the staff identified by

our Executive steering group. The questionnaire covered the areas listed below and was deliberately left broad enough so as to be easy to start the information flowing back to us.

Location, location, location

- Do you know where your digital assets are?
- If they are related to a physical (analogue) item, are they colocated with that item?
- If not, where are they?
- What conditions are they being stored in?

Retrieval

- Can we achieve easy access to them?
- Can they be sent to us?
- Are they catalogued?
- How many digital assets do you have?
- How big (MB) are the digital assets?
- Is their number of assets considered large? i.e. will we have to examine only a sample set?

Identification

- What media formats do you have?
- (CD (ROM, R, RW, Audio), floppy (various kinds), hard disk (IDE, SCSI, ESDI, etc.), magnetic tape
- What file formats do you have?
- What software environment (operating system, applications) is required to use the assets in question?
- What hardware environment is required to use the assets in question?
- Is there material that you know you have already lost access to?
- Is there material that you would deem to be at high risk?

The questionnaire was well received and alongside our internal communications and reporting structures formed the communication plan.

Establish the context

Once the communication plan was set-up and approval to proceed had been assured, the context of the study had to be drawn up. As stated in the introduction there needs to be a clear relationship between what you are trying to achieve and your corporate or institutional strategies. For the BL this was a matter of looking through our various strategic documents to find the correct measures of value to place our risk assessment with.

The BL follows a number of important legislative and strategic documents. The DPT split this responsibility into an internal (to the BL) and external (to the BL) context.

Internal context

The British Library has clearly outlined its commitment to safeguarding digital objects and to making these objects accessible. The Library's 2005-2008 strategy highlights the following points as critical to the ongoing purpose, goals and objectives responsibilities of the organisation:

The British Library Strategy 2005-2008:

- Strategic priority 1: Enrich the user's experience
- Strategic priority 2: Build the digital research environment
- Strategic priority 3: Transform search and navigation
- Strategic priority 4: Grow and manage the national collection

Other relevant BL strategies

- E-IS strategy (the BL's IT strategy)
- S&C content strategy (the BL's Collection strategy)
- 10 Year Digital Preservation strategy

Additionally, the Legal Deposit Libraries Act 2003 and the Irish Copyright Act 1963 (currently being replaced by similar provisions in the Copyright and Related Rights Bill 1999) place upon The British Library the responsibility to maintain legal deposit publications. These publications can include digital objects and, although not expressly covered under existing legislation, the stewardship of these objects must be considered. A proposed extension to legal deposit to cover digital objects is pending and is expected to pass sooner rather than later, so is included here as a contextual basis to be considered.

Within The British Library, there are a number of strategies that also add to the context. The e-IS strategy and the digital preservation strategy both set out clearly the responsibilities for effective stewardship of digital objects.

Ensuring the long-term accessibility of digital assets is the goal of the Digital Preservation Team. There are a number of tiers of accessibility, with each higher tier dependent on the lower tiers. Specifically:

- Bit-stream preservation: The raw sequence of bits stored on a digital medium must be readable. This requires safeguarding of digital media and/or migration to more robust media as necessary.
- File preservation: The bits must be interpretable as a usable digital object; this means developing or preserving suitable software/hardware to open the file, or performing migrations on the file, or some combination thereof.
- Semantic preservation: The files themselves typically constitute part of a greater whole (for example, each file may represent a scanned page of a book), and to be given meaning (for example, "this is page X of book Y") requires the creation and preservation of suitable metadata. Similarly, suitable metadata must exist to allow retrieval and discovery of the objects in the first place.

In keeping with these strategic responsibilities, the recommendations from the risk assessment were able to take the form of

- Technical recommendations (what to do with the material we already have to safeguard it)

- Organizational/Procedural/policy recommendations (to cover all stages of the lifecycle, from ingest through to long-term storage and preservation)
- Acquisition recommendations (given the choice the Library would prefer to acquire low-risk items)

R02	Physical damage	Medium	General
R03	Environmental Damage	Medium	General
R04			General

External context

The methodology defines this section as addressing the business, social, regulatory, cultural, competitive, financial, and political demands placed on organization.

External stakeholders to The British Library include Department of Culture Media and Sport our parent body. In redefining the library⁴ the annual report for 2005/06 the library outlines its responsibilities these include

- Responsible to Department of Culture Media and Sport
- Other UK legal deposit libraries
- Research Community/Higher Education
- General Public

These external stakeholders expect accountability for the safekeeping of all library assets, part of this is the management of digital objects within our collections, this risk assessment goes some way to illuminating how this management can be done in a digital environment

R04a	Technical Obsolescence	Medium	CD-ROM/CD-R/CD-RW
R04b			DVD-ROM/DVD-R/DVD+R/DVD-RW/DVD+RW
R04c			Floppy disk (e.g. 8", 5.25")
R04d			Floppy disk (3.5")
R04e			Hard disk
R04f			Magnetic tape (e.g. IBM 3480)
R04g			Magnetic tape (e.g. LTO3)
R04h			Other magnetic media
R04i			Paper tape/punch card
R04j			Other
R05			General

Identify the risks

The e-collections analysis has identified a number risks to digital objects across the collections. This enables us to group together the common themes and pull out the risks in groupings in order to rank them.

The identification and analysis of the collection area material has given us around 23 numbered risks, these risks are numbered from R01 through to R023. These risks once can usefully be grouped into 23 key risks to the collections. These 23 risks are as follows

R05a	Technical Obsolescence	File system	FAT
R05b			NTFS
R05c			HFS
R05d			ISO 9660
R05e			UDF
R05f			ADFS
R05g			OFS
R05h			FFS
R05i			(Other obsolete/legacy file system)
R06			

Reference	Risk	Type	Subtype
R01			General
R01a			CD-ROM
R01b			CD-R
R01c			CD-RW
R01d			DVD-ROM
R01e	Physical deterioration	Medium	DVD-R/DVD+R
R01f			DVD-RW/DVD+RW/DVD-RAM
R01g			Floppy disk
R01h			Hard disk (online)
R01i			Hard disk (array)
R01j			Hard disk (offline)
R01k			Magnetic tape (e.g. IBM 3480)
R01l			Magnetic tape (e.g. LTO3)
R01m			Paper tape/punch card
R01n			

R06a	Technical Obsolescence	File format	JPEG		
R06b			GIF		
R06c			TIFF		
R06d			JPEG 2000		
R06e			Broadcast Wave		
R06f			NTF		
R06g			Word .doc		
R06h			Excel .xls		
R06i			Photoshop .psd		
R06j			Wordstar (etc.; legacy software)		
R06k			"Programs"		
R07			General		
R07a	Technical Obsolescence	Hardwar environ.	PC		
R07b			Amiga		
R07c			Atari		
R07d			Acorn		
R07e			Apple Mac		
R07f			Sun		
R07g			Other		
R07h					
R07i					

R08			General
R08a			
R08b			
R08c			
R08d			DOS
R08e			Windows 3.x
R08f			Windows 9x
R08g			Windows NT
R08h			Windows 2000/XP
R08i			Windows XP non-Latin
R08j	Technical Obsolescence	Software environs	MacOS X
R08k			MacOS 9/below
R08l			AmigaOS
R08m			Atari TOS
R08n			Acorn RISC OS
R08o			Linux
R08p			Solaris
R08q			Niche obsolete operating system
R08r			Word
R08s			Excel
R08t			Acrobat Reader
R08u			Photoshop
R08v			NTF software
R08w			Broadcast Wave software
R08x			Wordstar (etc.; legacy software)
R09			Complex process for digital acquisitions that discourages material from being collected
	Acquisition		
R10			Insufficient up-front planning of storage and handling requirements
R11			No standardized verification of acquired media
R12	Ingest		No standardized analysis of acquired media
R13			No standard handling of acquired media
R14			Inadequate cataloguing of digital assets
R15	Metadata	Policy	Insufficient creation of metadata
R16			Limited usage statistics collected
R17			Little up-front consideration of who will access material and how they will do it
	Access		
R18			Internal IT policy causing premature loss of access
R19			DOM not ready to use
	Storage		
R20			Project-based funding does not always address storage
R21			Lack of digital curators
	Preservation		
R22			Lack of developed digital preservation tools
R23			Limited DPT resources

Analyse the risks

The 23 identified risks were then analysed using a combination of the DRAMBORA trusted repository

impact scale and industry analysis of the characteristics and the deterioration rates of physical media. Physical media all undergo a certain amount of deterioration naturally; even if kept in ideal circumstances, their lifetimes are finite due to unavoidable decay of their components. Media types are split into optical, magnetic and all others and the types of damage were identified as physical, and environmental. Additionally obsolescence of hardware and software, arguably the most pressing concerns from a digital preservation point of view were used to evaluate the risks at this point. Physical and environmental damage was useful to identify the people and organisational risks. This means that we are able to document and recommend future activity to reduce risk in this area alongside the technical obsolescence thereby covering not just what is at risk today but addressing what might be at risk tomorrow.

Evaluate the risks

At this point we Evaluate and compare to the organisations objectives. This evaluation has allowed us to compare using our LIFE⁵ methodology, the procedural and organisational gaps that have enabled us to plan for future work. Using a lifecycle methodology we are able to track the digital objects whether they are CDROM or DVD and use the methodology to streamline or make recommendations to tighten existing systems. This follow-up work is called the Acquisition and Handling study which will focus in part on training needs and system requirements to reduce the overall risk of the collections. They are;

Creation

- The digitisation approvals process does not cover all projects within the BL. Many projects are still co-ordinated from the Business areas of the BL. (currently now being addressed)

Acquisition

- Inadequate planning and consideration of what to do with large-scale digitisation output—nowhere centrally to put acquired content.
- There is not enough up-front consideration of digital preservation needs.

Ingest

- No standard verification of received media. No standard analysis of received media (i.e. the specific nature of the digital acquisitions is unknown)
- No standard handling/storage of received media. In most cases, the digital object is treated as a lesser priority, with the result that many digital objects are stored in suboptimal conditions.

Metadata

- No standard cataloguing of received media, meaning that there is no real understanding or knowledge of what it is we hold. (now being addressed)
- No BL standard (minimum implementable amount) of metadata for digital projects

- No comprehensive recording procedure of what disks have come from what source
- No extractions of available metadata, no tools on ingest to help.
- No good usage statistics are collated for digital objects.

Access

- Not considered at point of entry—who is the target market for the acquired material and how will they gain access.
- Some technical problems (especially software compatibility—unavailability of non-Latin Windows for example) are not ultimately technical (the software is widely and readily available) but can be policy.
- Some collection area content may only be accessible on previous versions of operating environments.

Storage

- DOM (now DLS) central storage is available but not ready to help with this. There is a need for a service to help mitigate the risks. (now underway)
- Project-based funding does not always address storage concerns.

Preservation

- Lack of widespread digital stewards within the collection areas
- Lack of developed tools and services to aid preservation. There is work being done in this area by the Planets⁶ project. However, there is still a time gap between this risk assessment and the end of the Planets project.
- DPT resources limit what we can do to help. A separate resource plan needs to be worked out so that the identified risks can be given a timetable for rescue.

This assessment of the policy issues surrounding the technical issues have brought to our attention the areas in most need of follow-up consultation

From this combination of media type, risk faced and policy and organisational objectives it is now possible to group the 23 risks into 8 categories and rank them in order to start to mitigate the risk faced by those most pressing.

Risk ranking	Risk	Access type jeopardized
8	Media degradation	Bit-stream
7	Media obsolescence	
6	File format obsolescence	File/Semantic
5	Hardware obsolescence	
4	Operating system + file system obsolescence	

3	Software obsolescence	
2	Poor policy (improper cataloguing, metadata)	Semantic
1	Poor policy (other)	Semantic/File/Bit-stream

Treat the risks

In terms of the risk assessment itself, treating the risks was considered to be outside the scope. However it is very important to note that the treatment of the risks identified has formed the major part of a funding bid within the BL to address the needs identified. Up until this point it was thought to be the case that hand-held media had a shelf-life that was in keeping with the timeframes to ingest this material. It was actually the case that urgent action has had to be done and so treatment for the risks now falls to the DPT under the name of content stabilisation, this work is currently in progress and is expected to form a vital part in the overall National Digital Library Programme for the UK in coming years. The facility is now installed within the BL's centre for digital preservation and is currently conducting analysis of 120TB of digitised newspaper content.

Monitor and review

Risk assessment requires a continuous improvement approach to be effective. The document is a tool for digital preservation activity and has prioritised the most at risk parts of The British Library's digital collections. From this list, action can be taken to reduce the risk and to preserve the content in a continuous manner. In order to achieve this, the assessment will be re-evaluated each year.

The purpose of this re-evaluation will be to reduce the numbers in the prioritisation table, representing an overall reduction in risk to the collection. This performance will be monitored and reviewed by the Digital Preservation Team so reduction in risk will become a key performance indicator for the Digital Preservation Team.

The key performance indicator and prioritisation table will become the overriding driver for future digital preservation activity in the area of collection based electronic content. The Digital Preservation Team's activity in this area will provide a continuous assessment of technical obsolescence, the viability of format migration, and availability of emulation technology. This may result in changing priorities or the development new mitigation strategies, where these occur updates will be added to the prioritisation table.

From the prioritisation table it has been agreed by the Digital Preservation Team that all collection content identified as category 8 risk will be addressed first. In order to do this a resource plan will be created separately from this assessment document. This will outline the time, cost, and effort required to tackle all objects within

the highest risk category. If the cost is felt to be within the capabilities of the current Scholarship and Collections/Electronic Information Services budget the Digital Preservation Team will take the management of these risks to the next stage of mitigation, actively moving the data to a more stable environment. At this point, the resource plan will become part of the monitoring process.

Summary of monitoring action points:

- Annual update to the risk assessment to continuously improve the condition of the collection based digital objects.
- Annual identification of resulting actions to mitigate risks.
- Management of the digital preservation prioritisation table.
- Key performance indicators to be drawn from the risk factors within the prioritisation table, to be monitored by the digital preservation steering group. (Ideally all risk factors should be in a continuous process of reduction).
- Business change functions are being monitored.

Concluding statement

Digital Preservation is much more than a technological problem, its management and measurement requires it to be embedded throughout any organisation. The risk assessment carried out at the BL could be expected to return a similar result regardless of where an institution is geographically or what the organisations function is. The common denominators to this work are what value is placed upon the digital content and what resource is available to do something about it. The ability to use risk as a catalyst for change is a powerful argument and one which has proven beneficial to not just our understanding of the content but our understanding of the organisation and the policies that govern its existence. It is expected by the BL DPT that this work and its subsequent follow on exercises in Acquisition and Handling and content stabilisation will form an important part in our future efforts to preserve e-collections.

British Library digital preservation strategy
www.bl.uk/dp

Digital Repository Audit Method Based on Risk Assessment
<http://www.repositoryaudit.eu/>

Redefining the Library: The British Library's strategy 2005-2008
<http://www.bl.uk/about/strategy.html>

McLeod, R and Wheatley, P and Ayris, P. (2006) Lifecycle information for e-literature: full report from the LIFE project. Research report. LIFE project, London, UK. 122p. <http://eprints.ucl.ac.uk/archive/00001854/>

Redefining the Library: The British Library's strategy 2005-2008
<http://www.bl.uk/about/strategy.html>

Kings of all we survey
<http://www.guardian.co.uk/saturday/story/0,,2041775,00.html>

British Library's Content Strategy: Meeting the Needs of the Nation
<http://www.bl.uk/about/strategic/pdf/contentstrategy.pdf>

Longevity of CD media

<http://www.loc.gov/preserv/studyofCDlongevity.pdf>

Stability Comparison of Recordable Optical Discs
<http://nvl.nist.gov/pub/nistpubs/jres/109/5/j95sla.pdf>

Longevity of high density magnetic media
<http://www.thic.org/pdf/Nov02/nara.vnavale.021106.pdf>

Disk failures in the real world
<http://www.usenix.org/events/fast07/tech/schroeder/schroeder.html/index.html>

Failure Trends in a Large Disk Drive Population
http://209.85.163.132/papers/disk_failures.pdf

Planets project www.planets-project.eu

The Significance of Storage in the 'Cost of Risk' of Digital Preservation

Richard Wright*, Matthew Addis**, Ant Miller*

* BBC Research and Innovation
Broadcast Centre
201 Wood Lane
London W12 7TP, UK
richard.wright@bbc.co.uk;
ant.miller@bbc.co.uk

**IT Innovation Centre
University of Southampton
2 Venture Road
Southampton SO16 7NP, UK
mja@it-innovation.soton.ac.uk

Abstract

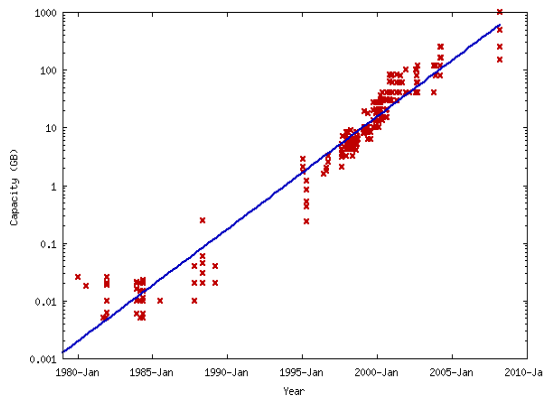
As storage costs drop, storage is becoming the lowest cost in a digital repository – and the biggest risk. We examine current modelling of costs and risks in digital preservation, concentrating on the Total Cost of Risk when using digital storage systems for preserving audiovisual material. We present a managed approach to preservation, and the vital role of storage and show how planning for long-term preservation of data should consider the risks involved in using digital storage technology. Gaps in information necessary for accurate modeling – and planning – are presented. We call for new functionality to support recovery of files with errors, to eliminate the all-or-nothing approach of current IT systems, reduce the impact of failures of digital storage technology and *mitigate against loss* of digital data.

Significance of Storage

As storage costs continue to drop by roughly 50% every 18 months, there are two effects:

- *Storage looks free (but isn't)*: the cost of storage devices becomes negligible, but power, space, cooling, management and replaced costs remain significant.
- *Storage is abundant*: much more storage is used

The following figure shows how hard drive storage has increased over the last 25 years (Hankwang 2008).



The largest available size (for a desktop computer) has increased from 5 MB to one terabyte – a factor of 200 000 (which is about 18 doublings in about 25 years, so very close to doubling every 18 months).

The 'growth of risk' is of course much larger: a factor of 200 000 in disc size, times the increase in the usage of discs (about 10 000 over the same period; Computer World, 2008).

This "growth of storage" also divides into two effects:

- the number of storage units (globally, and used by any given institution) increases
- the amount of data stored on each unit also increases

The increase in storage units (devices) means that statistics on failure rates that were once seen as 'safe' are now appreciable risks. An advertised Mean Time Between Failure of 1000 years looks very safe to a person buying a new hard drive (though it will be obsolete in 5 years). Schroeder and Gibson (2007) give results on a survey of major datacentres holding 100 000 discs, and found annual failure rates ranging from one to 13 %, averaging around 3% - far higher than an MTBF of 1000 years.

This failure rate means that owners of 1000 of those same hard drives will need systems (eg big RAID6 arrays) and processes (eg continual hot-swapping and rebuilding) to ensure these failures are managed..

The increase in storage units results in more and more users being responsible for, or dependent upon, storage systems that have thousands of individual storage devices (hard drives, optical discs, data tapes). The increase in the amount of data stored on each device makes the failure of each device more significant in terms of the volume of data potentially lost. A 3.5" floppy disc with 1.4 megabytes (MB) of data represented a few dozen files. A 650 MB CD could hold 500 times more data: thousands of files, or one hour of audio. A USB-attached terabyte hard drive is 700 000 times

Comment [mja1]: 'Storage is free' is a dangerous statement to make – if storage is free then keeping multiple copies is free and hence there is no cost in reducing risk – you can do LOCKSS for free.

bigger than a floppy, and 1400 times bigger than a CD. It could, for example, hold the entire contents of an institution's audio collection (such as several years' work by many people, collecting oral history recordings).

Cost Modelling

We will present an approach to risk that combines the dimensions of cost, risk (uncertainty) and value (benefits). This model builds upon and extends work on cost modelling by both the digital library and audiovisual communities. Early on in the development of digital libraries there was the fundamental work on preservation strategies by Beagrie and Greenstein (1998), Hendley (1998), Granger, Russell and Weinberger (2000) – and eventually something about the audiovisual sector from EU PRESTO project (Wright, 2002). The state of the art was brought together, and specifically labelled 'life cycle', in the important paper of Shenton (2003).

Since then, there have been entire projects and conferences devoted to *life-cycle models and costs*. At a conference organised by the Digital Preservation Coalition and the Digital Curation Centre (DPC/DCC 2005) there were reports from the LIFE and eSPIDA projects, both specifically about costs, though the eSPIDA work was more generally concerned with a formal method for including intangible benefits (value) in business cases. More pertinent to the present paper, it also specifically introduced the issue of uncertainty into the modelling process.

Specific digital library and digital preservation cost models reported at the 2005 DPC/DCC conference included work from Cornell University, TNA in the UK, and the Koninklijke Bibliotheek in the Netherlands as well as two papers arising from PrestoSpace. In all these models and studies, and for digital library technology in general, little is said about storage (except in the PrestoSpace work). Digital libraries assume that storage will be there (somewhere), and will work and continue to work. In estimating Total Cost of Ownership (TCO), the complexity of the models just mentioned is devoted to digital library processes, not storage devices (or their management). In digital library/repository TCO models, storage cost is generally modelled as a single number per year, and the model simply 'adds up' those numbers.

Cost-of-Risk Modelling

Estimation of cost involves uncertainties. Some uncertainties can be represented as variances in cost estimates (uncertainty about how much costs may vary from the predicted value), but a whole range of uncertainties are related to things that may or may not happen, and should be formally identified as **risks**.

A risk is the likelihood of an incident along with the business consequences (positive or negative) (Addis, 2008a).

Examples of possible incidents include:

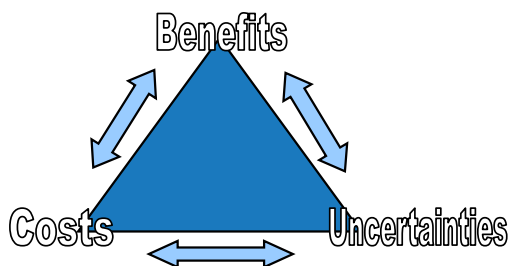
- Technical obsolescence, e.g. formats and players
- Hardware failures, e.g. digital storage systems
- Loss of staff, e.g. skilled transfer operators
- Insufficient budget, e.g. digitisation too expensive
- Accidental loss, e.g. human error during QC
- Stakeholder changes, e.g. preservation no longer a priority
- Underestimation of resources or effort
- Fire, flood, meteors ...

Traditional risk modeling (and its use in project management) looks at lists of such incidents, and their attendant likelihoods (assessing likelihood may have the largest uncertainty of the whole process!) as contained in a risk register, and then proceeds to predict the consequences – the impact – of each item.

Possible consequences for preservation from the above list of incidents would include:

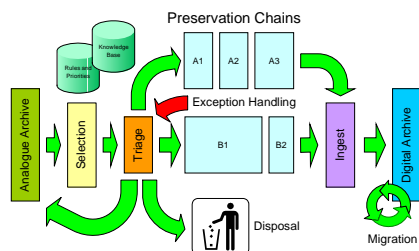
- Corruption or loss of audiovisual content
- Interruption to services
- Inefficiencies and increased costs
- Corner cutting and increased risks
- Failure to meet legal obligations
- Loss of reputation or loss of customers

A more comprehensive approach to the whole issue of uncertainty in preservation is to include the concept of value (benefit). The work of eSPIDA has already been mentioned.



The combination of uncertainty, cost and benefits forms a three-way interaction, as shown in the above diagram. The key point about this approach is that it is applicable to the whole issue of business-case planning, not just to the more narrow issues of risk analysis and cost modeling.

A typical preservation scenario, which can be optimized by use of the cost-of-risk approach, is given in the following diagramme:



This integrated approach to cost, risk and value allows all the factors affecting preservation planning, funding and management to be considered in one set of interactions, rather than being taken separately.

For quantitative modeling, all three factors need to be converted to a common unit of measurement. As cost and benefits are already commonly thought of in financial terms, the task is then to also express the uncertainties in monetary units: the cost-of-risk.

Full details require a much longer presentation. There has already been a great deal of detailed work, specifically relevant to preservation, in the DRAMBORA project (DRAMBORA 2008), and much more detail is in Addis (2008a).

The following diagram shows the consideration of risk as the central metaphor in strategic planning.



Minimisation of Risk and Cost of Risk – and Mitigation of Loss

The effort within the digital library community to define and construct trusted digital repositories pays little attention to storage. The *trust* issue is defined and examined mainly at the institutional level, not at the level of IT systems and certainly not at the level of individual device or file failures. Yet the *only* physical reality of the content of a trusted digital repository lies in its files, sitting on its storage. The 'atomic level' of success or failure of a repository is the success or failure of an attempt to read individual files. Such success or failure is clearly fundamental to the concept of trust for the whole repository.

Effort of the storage area of the IT industry is entirely focused on reducing the likelihood of read errors (device failure or file read error). There is no concept, within

standard IT systems, of a partially-recoverable file. If the inevitable low-level errors cannot be corrected by the built-in error detection and correction technology, the read fails and the file fails to open. There is nothing that the ordinary user can do at this point, and even the all-powerful system manager can only look at backups to see if there is another copy of exactly the same file. There is technology to attempt to read corrupted files or failed hard drives, but such technology falls in the category of *heroic measures*: sending the file or drive to an external company that will attempt a recovery using proprietary technology, at a substantial price (see reference: Recovery Tool Box).

Physically, a file with a read error is not an all-or-nothing situation. There will still be a stream of data (somewhere in the *stack* of operations between the user and the hardware) which is likely to be mainly correct, and is also likely to even have indications of which bytes are incorrect (because of lateral parity errors). For simple error detection and correction schemes, a common situation underlying an inability to read a file is a single block of data that has two or more such errors, so that the longitudinal parity check is ambiguous. At that point, a whole file of many blocks of data is called unreadable, because two bytes – at known locations – fail their parity check and so are known to be erroneous.

Returning to the definition of risk as having two factors: *probability* and *impact*: the ability to read *most* of the data in a corrupted file would, in certain cases, greatly reduce the *impact* of the error. This is the area of risk reduction that is being examined by the UK project AVATAR (Addis et al, 2008b; AVATAR is also looking at the whole issue of optimization and management of storage, from the perspective of archiving and long-term preservation).

Reducing the impact of a storage failure is a method for *mitigation of loss* (Knight, 2007). The issue of loss and recovery from loss has been identified as a neglected area in digital preservation thinking, but its importance has been highlighted by the growing awareness of the phenomenon of bit rot (see reference).

Despite the best efforts of the IT industry, despite mean time between failure of hard drives exceeding one million hours, and despite tests of storage functionality yielding read-error estimations of one failure in 10^{17} read attempts – errors do occur. The author has, in 2008, been personally experiencing one file read failure per month – and in each case these are total failures, with no possibility of mitigation (beyond the commercial route of heroic measures).

Redundancy and Risk

Standard practice for reducing risk of loss is to have another copy. The use of second (or higher) copies is a method of reducing impact: a file read error or a device failure has much less impact if recourse can be made to a backup copy or system.

At a more sophisticated level, arrays of hard drives are used to gain the benefits of redundancy at lower cost. RAID (see reference) technology achieves protection for the loss of one of N drives in a set of $N+1$ – so the net cost is $N+1$ drives, rather than the $2N$ that would be required by simple redundancy.

RAID has now advanced (e.g. RAID6) to the point where multiple disks can fail without data loss, which means data can still be accessed safely whilst individual disks are being replaced and live rebuilding takes place. This allows disk systems to be built that are resilient to hardware failures, human errors and data read errors. For large data centres, the problem is shifted from risk of loss from device failure to having the right support processes to ‘feed’ large systems with a constant supply of new drives and have the people in place to do so.

At the same time as redundancy is added to storage systems to reduce risk, redundancy is being taken out of the files stored on those systems, as a way to save space. Compression, lossless or lossy, is based on the innate redundancy (entropy) of the original data. When the redundancy is removed from a file, a complex transformation has to be applied to the resulting data in order to transform it back to the original (or close to the original, in the case of lossy compression).

To Encode or Not to Encode

The process of compressing (encoding) a file has profound consequences for attempts to mitigate against loss. A consequence of removal of redundancy is that the remaining data is all very significant – because a compression process is entirely an attempt to eliminate insignificant data. If one byte of the resultant file is then damaged, that byte is then very likely to be used involved in computations (the decoding or decompressing process) that will affect many other bytes. Encoding a file severely affects the ability to use corrupted data as a method of reducing the impact of error.

As an example: an uncompressed audio .WAV file is simply a header followed by a sequence of numbers – one number per sample of the desired audio waveform. If the audio is sampled at 44.1 kHz (the rate used on CDs), each sample represents about 23 micro-seconds of data. Losing one byte of data results in one bad sample, but there is no spread to any of the rest of the data.

Hence an uncompressed audio file can be perfectly usable despite loss of one byte. Indeed, experiments have shown that a .WAV file with 0.4% errors is almost undistinguishable from the original, whereas an MP3 file with the same level of errors either will not open at all, or will have errors affecting most of the audio, and rendering it unusable.

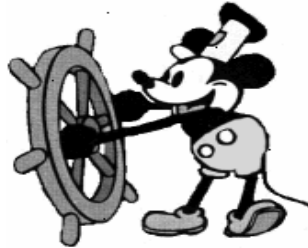
The same logic applies to video, images – and even to text if represented as a sequence of characters (with embedded mark-up, as in the old days of ‘printer control characters’ as escape sequences within a text ‘stream’).

An extensive study of the consequences of byte-level errors on different file types, compressed and uncompressed, was recently presented by Heydegger (2008). His results include the following data for image files; in each case exactly one byte had been changed:

- o a 10 MB TIFF = .000 01% errors (meaning just that one byte affected)
- o a lossless JP2 had 17% errors for a saving of 27% in storage
- o a lossy JPEG had 2.1% for a saving of 62% in storage

Comment [mja2]: Is this really 17% for a saving of 27% storage? - that's a pretty crap trade off.

As an example of the affect of data loss on imager files, here are two examples: a BMP (uncompressed) and a GIF (compressed). Each had one byte in 4k changed – meaning 3 bytes total for the GIF, and 12 for the BMP



Not quite ready for retirement:
Steamboat Willie, 1928

Used without permission
BMP with one error every 4K bytes

GIF file with one error every 4K bytes.

From the above results, it is evident that removing redundancy increases impact, the “cost of error”. The compression increases the proportional damage caused by an unrecoverable read error. However if there is no mechanism for using files despite read errors, then it is of no practical significance whether a one-byte error causes major damage, or only very local and very minor damage. If the file can’t be read in either case, the error-magnification factor caused by compression is hidden.

If less-than-perfect files can be passed back to the user, or to a file restoration application, then the increase in

'cost of error' caused by compression can be legitimately compared with the decrease in cost of storage.

An unsolved issue in preservation strategy is whether it is better (lower 'cost of risk' for the same or less total risk) to use lossless compression and then make multiple copies (externalized redundancy) as a way to reduce the impact of storage errors – or to avoid compression and exploit the internal redundancy of the files. The problem at present is that there is little or no technology (within conventional storage systems, or conventional digital repositories) to support the second option.

The question of which strategy to take depends on more than just the ability of file systems to return files with partial errors. A holistic approach to risk management means dealing with disaster recovery (fire, flood, theft etc.), human error (accidental corruption, deletion, miscataloguing etc.), and technology obsolescence (formats, software, devices etc.). All present powerful drivers for multiple copies in multiple places using multiple technical solutions. If an offsite copy of uncompressed video is created to address DR, then lossless compression may allow two offsite copies for the same cost. Three copies in three places may well be enough to reduce the risk of loss due to individual storage failures to a level where no further measures are needed beyond those of conventional storage systems, e.g. RAID.

However, until file reading systems are willing and able to return files despite errors, and include media-specific reconstruction techniques to 'fill in' where errors are known to exist, there will be no effective way to exploit file-error recovery as a method to mitigate against loss. This prevents a whole class of 'cost of risk' strategies from being used to complement conventional techniques.

The frustration for audiovisual archivists is that digital technology has taken us one step forward, and now is taking us two steps back. The ability of analogue videotape recorders to cope with loss of data (dropout) was limited, and black lines would appear in the resultant images. Digital tape recorders had much better built-in compensation: the *concealment* option would allow a missing line to be replaced by a neighbouring line, and expensive machines could even replace entire frames with an adjacent (in time) frame. Now file-based digital technology has no ability to cope with loss, beyond the 'external redundancy' option of multiple copies.

One could accept that files are, and will remain, 'all or nothing' entities – you either get everything in them or you lose the lot. The strategy then becomes one of splitting assets, e.g. a video sequence, into multiple files and then implementing safety measures at the 'application' level. For example, an audiovisual program could be split into separate files for shots, scenes, frames, regions of interest, audio, video or many other ways. The most important parts would then be assigned to one or more storage systems with appropriate levels of reliability – avoiding the 'all eggs in one basket' problem. The advantage here is that how to 'split' an

asset into pieces can be done based on an understanding of what the asset is – something that a file system or storage device will never have. The downside is increased technology and management costs – a violation of the 'simplest is best' principle.

We hope that current work in preservation theory and methodology, with use of file description metadata, will support and encourage the ability of storage systems to return less-than-perfect files in a usable fashion.

Examples of work with relevance to file description include Planets (file characterization) and Shaman:

- MPEG-21 DIDL = Digital Item Declaration Language (see File Description reference)
- XCEL, XCDL = eXtensible Characterisation Languages (Becker, 2008; Thaller, 2008)
- Shaman = multivalent approach (Watry, 2007)

Conclusions

Comprehensive and integrated planning for preservation can be accomplished through use of a three-factor model, based on costs, benefits and uncertainties. The cost-of-risk concept allows all three factors to be quantified on a common, monetary scale.

Reduction of the cost-of-risk, and the best chance for mitigation of loss, is by **always taking the simplest option** – beginning with not compressing the data.

Storing only uncompressed data would appear to add cost rather than reduce it – but storage costs are typically a small part of a preservation project or strategy (labour is always the dominant cost), and storage cost is dropping by 50% every 18 months.

The full benefit of uncompressed files (in terms of mitigation of loss and consequent reduction of impact) will remain irrelevant unless and until the storage industry and digital repository architects produce systems that allow access to less-than-perfect files.

References

Addis, M (2008a) Cost Models and Business Cases (for Audiovisual Curation and Preservation) presentation at HATII-TAPE Audiovisual Preservation Course, Univ of Edinburgh, May 2008.
www.hatii.arts.gla.ac.uk/news/tape.html

Addis, M et al, (2008b) Sustainable Archiving and Storage Management of Audiovisual Digital Assets; to be presented at IBC 2008, September, Amsterdam.

AVATAR-m:
<http://www.it-innovation.soton.ac.uk/projects/avatar-m/>

Beagrie, N & Greenstein, D (1998) A strategic policy framework for creating and preserving digital

Comment [mja3]: I deleted this paragraph as I don't think it's true. Device costs (disks or tapes) may become negligible, but TCO for storage (inc. maintenance, upgrade, power, space, people etc.) certainly won't – see intro section.

collections. British Library Research and Innovation Report 107. London: British Library.

www.ukoln.ac.uk/services/elib/papers/supporting/pdf/framework.pdf

Becker, C et al (2008) A generic XML language for characterising objects to support digital preservation. Proceedings of the 2008 ACM symposium on Applied computing, Fortaleza, Brazil. pp 402-406
<http://portal.acm.org/citation.cfm?id=1363786&jmp=ind&extms&coll=GUIDE&dl=GUIDE>

Bit Rot: http://en.wikipedia.org/wiki/Bit_rot

Chapman, S (2003) Counting the Costs of Digital Preservation: Is Repository Storage Affordable? Journal of Digital Information, Volume 4 Issue 2 Article No. 178, 2003-05-07

Computer World (2008) Seagate ships one-billionth hard drive.

http://www.computerworld.com/action/article.do?command=viewArticleBasic&taxonomyName=storage&articleId=9079718&taxonomyId=19&intsrc=kc_top

DPC/DCC (2005) Workshop on Cost Models for preserving digital assets
www.dpconline.org/graphics/events/050726workshop.html

DRAMBORA (2008) www.repositoryaudit.eu/

eSPIDA: (costs, benefits and uncertainties in project proposals) www.gla.ac.uk/espida/

File Description methodologies:

DIDL: <http://xml.coverpages.org/mpeg21-didl.html>

XCEL, XCDL: see Becker, 2008 and Thaller, 2008

Granger, Russell and Weinberger (2000) Cost models and cost elements for digital preservation CEDARS project
www.leeds.ac.uk/cedars/colman/costElementsOfDP.doc

Hankwang (2008)

http://en.wikipedia.org/wiki/Image:Hard_drive_capacity_over_time.png

Hendley, T (1998) Comparison of methods and costs of digital preservation. British Library. Research and Innovation Report, 106
www.ukoln.ac.uk/services/elib/papers/tavistock/hendley/hendley.html

Heydegger, V (2008) Analysing the Impact of File Formats on Data Integrity. Proceedings of Archiving 2008, Bern, Switzerland, June 24-27; pp 50-55.

Knight, S (2007) Manager Innovation Centre and Programme Architect National Digital Heritage Archive, NLNZ. Remarks 'from the floor' on the significance of efforts to mitigate against loss, at the SUN PASIG meeting, Nov 2007, Paris.

http://sun-pasig.org/nov07_presentations.html

LIFE: Life Cycle Information for E-Literature
www.life.ac.uk/

Moore's Law: http://en.wikipedia.org/wiki/Moore's_law

Palm, J (2006) The Digital Black Hole:
www.tape-online.net/docs/Palm_Black_Hole.pdf

PrestoSpace: www.prestospace.eu
<http://digitalpreservation.ssl.co.uk/index.html>

RAID: Redundant Array of Inexpensive Discs – an efficient method of achieving device-level redundancy.
en.wikipedia.org/wiki/Redundant_array_of_independent_disks

Recovery Tool Box <http://www.recoverytoolbox.com/>
This company is just one of many offering tools that *may* be able to repair a corrupted file.

Risk definition and risk management:

JISC: www.jiscinfonet.ac.uk/InfoKits/risk-management/

PRINCE2: www.ogc.gov.uk/methods/prince_2.asp

Shenton, H (2003) Life Cycle Collection Management LIBER Quarterly, 13(3/4)

Schroeder, Bi and Gibson, G A (2007) Disk failures in the real world: What does an MTTF of 1,000,000 hours mean to you? 5th USENIX Conference on File and Storage Technologies, Feb 2007
www.usenix.org/event/fast07/tech/schroeder/schroeder.html/

Thaller, M (2008) Characterisation (Planets presentation)
http://www.planets-project.eu/docs/presentations/manfred_thaller.pdf

Watry, P (2007) Digital Preservation Theory and Application: Transcontinental Persistent Archives Testbed Activity. International Journal of Digital Curation, 2-2. www.ijdc.net/ijdc/article/viewArticle/43/0

Wright, R. (2002) Broadcast archives: preserving the future. PRESTO project
http://presto.joanneum.ac.at/Public/ICHIM%20PRESTO%2028_05_01.pdf

Wright, R (2007) Annual Report on Preservation Issues
www.prestospace.org/project/deliverables/D22-8.pdf

International Study on Copyright and Digital Preservation

June M. Besek

Kernochan Center for Law,
Media and the Arts
Columbia Law School
jbesek@law.columbia.edu

Jessica Coates

ARC Centre of Excellence
for Creative Industries and
Innovation
Queensland University of
Technology
j2.coates@qut.edu.au

Brian Fitzgerald

Law Faculty
Queensland University of
Technology
bf.fitzgerald@qut.edu.au

William G. LeFurgy

Office of Strategic
Initiatives
Library of Congress
wlelf@loc.gov

Wilma Mossink

SURFfoundation
A.mossink@surf.nl

Adrienne Muir

Department of Information Science
Loughborough University
A.Muir@lboro.ac.uk

Christopher D. Weston

US Copyright Office
Library of Congress
cwes@loc.gov

Abstract

The International Study on the Impact of Copyright Law on Digital Preservation reviewed digital preservation activities and the current state of copyright and related laws and their impact on digital preservation in Australia, the Netherlands, the United Kingdom and the United States. In many cases, digital works are not being preserved in any systematic way, in part because digital preservation triggers copyright concerns in a way that analogue preservation does not. All the countries have some form of preservation exception. However, there is an inconsistent approach in the details and uncertainty as to how they may apply in the digital environment. None of the countries have a uniform national system collecting digital materials. Technological protection measures and private contracts may in some cases present significant practical barriers to preservation.. Current approaches to address these legal barriers are ad hoc and include approaching individual rights holders and some use of model licences. There are as yet no effective solutions to the issue of orphan works. Recommendations of the study include suggestions for drafting national policies and adapting laws with the aim of allowing preservation activities to be undertaken as necessary and in accordance with international best practice standards and to allow a uniform national system for the collection of digital materials by relevant state and national collecting institutions.

Background to the Study

The aims of this study were to: review the current state of copyright and related laws and their impact on digital preservation; to make recommendations for legislative reform and other solutions to ensure that libraries, archives and other preservation institutions can effectively preserve digital works and information in a manner consistent with international laws and norms of copyright and related rights; and to make recommendations for further study or activities to advance the recommendations.

The study partners were the Open Access to Knowledge Law Project, Faculty of Law, Queensland University of

Technology, Australia; the SurfFoundation in the Netherlands; the Joint Information Systems Committee in the UK and the US Library of Congress, National Digital Information Infrastructure and Preservation Program. Each of the study partners surveyed the situation in their own country and developed detailed country recommendations. The partners then developed a set of joint recommendations at a more general level for preservation institutions, legislators and policy makers. The partners also held a workshop in cooperation with the World Intellectual Property Organization in Geneva to present and discuss the study findings and recommendations, which have been published in a joint report (Besek, J.M. *et al*, 2008).

Digital Content, Digital Preservation and Copyright

Increasingly, radio and television programmes, musical compositions, films, maps, reports, stories, poems, letters, scholarly articles, newspapers and photographs are “born digital.” There is also a growing trend of converting so-called analogue material to digital form so that it can be easily and efficiently stored, transmitted and accessed. New forms of authorship, such as web sites, blogs and “user-generated content” of all kinds are flourishing in the dynamic environment of the Internet. These new works reflect the world’s culture as much as their analogue predecessors.

Embodying creative works in digital form has the unfortunate effect of potentially decreasing their usable lifespan. Digital information can be ephemeral: it is easily deleted, written over or corrupted. Because information technology such as hardware, software and digital object formats evolves so rapidly, it can be

difficult to access and use digital materials created only a few years ago.

Preservation is critical in the digital context to ensure continued long term access to historically, scientifically and socially valuable materials, so that future generations will be able to benefit from works created now. Libraries, archives and other preservation institutions have been responsible for much of the preservation that has occurred in the past. It is clear, however, that in many cases the digital equivalents of those works preserved in the past are not being preserved in any systematic way, in part because digital preservation triggers copyright concerns in a way that analogue preservation does not.

Digital preservation refers broadly to the series of managed activities necessary to ensure continued access to digital materials for as long as necessary, such as collection, description, migration and redundant storage. Digital preservation activities are undertaken by a range of preservation institutions, including libraries, archives and museums. Such institutions may operate independently or may be located within other bodies such as educational institutions, government entities or media organisations.

All of the countries discussed in this paper are members of the Berne Convention for the Protection of Literary and Artistic Works, which provides the foundation for governance of copyright law internationally. In addition, all have joined, or have indicated that they intend to join, the treaties that provide the principal modern updates to the Berne Convention – the World Intellectual Property Organization (WIPO) Copyright Treaty (WCT) and the WIPO Performances and Phonograms Treaty (WPPT), as well as the World Trade Organization’s Agreement on Trade-Related Aspects of Intellectual Property Rights (TRIPs). Together, these agreements require members to provide authors of literary and artistic works with a number of exclusive rights with respect to their works, including the rights of reproduction, adaptation, broadcasting, public performance, communication to the public and distribution to the public, subject to certain limitations and exceptions. In addition, performers of phonograms (also referred to in this paper as sound recordings) are provided with a right of fixation, and performers and producers of phonograms are granted rights of reproduction, distribution, rental, and making available their fixed performances. All of these rights are subject to limitations and exceptions.

Many of the activities involved in digital preservation, such as making multiple copies of a work, distributing copies among multiple institutions, and migrating works to new technological formats and media, involve the exercise of exclusive rights. For example, reproduction is a fundamental activity of digital preservation. The right of distribution may be implicated by disseminating digital copies to multiple institutions to protect against

catastrophic loss. And, to the extent access is required for digital preservation best practices, that access may implicate the right of “making available,” or of public performance or display. In any case access to content – either by users or by institution staff to verify its integrity – may entail making a copy on a screen and in computer memory. Other rights that may need to be considered are performance rights, rights in typographical arrangements, database rights (in European Union Member States, for example) and the moral rights of authors and creators.

Digital technologies have also changed the manner in which works are distributed and acquired in ways that create tension between long term preservation needs and copyright laws. Previously, copyright works were marketed in tangible “hard copy” form, and libraries, archives and other preservation institutions could acquire them on the market (or, in some cases, pursuant to legal deposit laws) for current use and long term preservation. But now, many works are never produced in hard copy. Some works – such as web sites and various types of “user-generated content” available on the Internet – are not made available for acquisition, but only for listening or viewing. Those works cannot be preserved by a preservation institution unless they can be copied or otherwise acquired. Other types of works such as e-journals are available on the market, but the terms of use may not permit the creation or retention of archival copies.

The unauthorised exercise of the rights in a work may result in infringement of copyright under the law of the various jurisdictions unless:

- the material is not protected by copyright (i.e., it is in the public domain);
- digital preservation is undertaken by the owner of copyright in the work or with the permission of the owner; or
- the copying or other use is permitted under an exception in the copyright law or related legislation (e.g., pursuant to an exception for libraries, archives or other preservation institutions or legal deposit).

The Berne Convention for the Protection of Literary and Artistic Works (1886) allows exceptions to the right of reproduction under certain conditions, known as the “three-step test”: member countries may permit limited copying of literary and artistic works for certain purposes so long as it “does not conflict with a normal exploitation of the work and does not unreasonably prejudice the legitimate interests of the author” (Art. 9(6)). The WIPO Copyright Treaty builds upon Berne’s three-step test by providing that contracting parties may allow limitations or exceptions to the rights granted under that treaty or under the Berne Convention in “certain special cases that do not conflict with a normal exploitation of the work and do not unreasonably prejudice the legitimate interests of the author.” (Art. 10). In other words, the

WIPO Copyright Treaty makes the three-step test applicable to exceptions and limitations with respect to any of the rights granted to authors under either that Treaty or the Berne Convention. The WPPT similarly makes the three-step test applicable to rights granted under that treaty. Thus, while these treaties do not mandate any exceptions or limitations specific to preservation activities or preservation institutions, the treaties do permit such exceptions or limitations, provided they comport with the three-step test.

The EU Information Society Directive (Directive 2001/29/EC of the European Parliament and of the Council of 22 May 2001 on the harmonization of certain aspects of copyright and related rights in the information society) permits, but does not require, European Union Member States to provide exceptions and limitations for certain activities of publicly accessible libraries, educational establishments or museums, or by archives. The permitted exceptions and limitations are: (1) for specific acts of reproduction of copyrighted works which are not for direct or indirect economic or commercial advantage, art. 5(2)(c); and (2) for use by communication or making available of copyrighted works in their collections, for the purpose of research or private study, to individual members of the public by dedicated terminals on the premises of such establishments, provided those works are not subject to purchase or licensing terms to the contrary, art. 5(3)(n).

Study Findings

Preservation and Other Relevant Exceptions

The four countries surveyed all have exceptions in their copyright and related laws that allow reproduction (and sometimes other activities) in connection with the preservation of protected works. However, many of the exceptions were enacted in an analogue era and do not adequately accommodate all of the activities necessary for *digital* preservation. The existing exceptions for preservation apply inconsistently across the jurisdictions with regard to which institutions may make use of them, the materials they apply to, the degree of copying they allow, and whether and how preservation copies may be accessed by the public.

For example in the UK, the exception refers to making “a copy”, where multiple and serial copying over time may well be required. Preservation exceptions may not apply to recorded sound or moving images. Exceptions may only refer to not for profit libraries and archives, so museums are excluded. Some countries have begun the process of changing their laws to create exceptions to allow digital preservation by libraries, archives and other preservation institutions, but applying the preservation exceptions that currently exist to digital preservation is often an uncertain and frustrating exercise. For example,

in the USA, libraries and archives exceptions allow only up to three copies for preservation and replacement, which is inadequate to maintain works in the digital environment.

None of the countries surveyed have provisions for so-called orphan works. These are works whose right holders cannot be identified or located. This is a particular issue for audiovisual material, photographs and illustrations. For example, the most effective method of preserving fragile recordings or films may be to digitise them. If the preservation exception does not extend to these types of material, preserving institutions will have to seek permission from rights owners to preserve them. If the rights owners cannot be traced, the material cannot be digitised, unless there is some other provisions, such as a fair use exception. The Australian and UK fair dealing exceptions only allow limited copying for a narrow range of purposes or news reporting. In the United States, libraries and archives are relying on the more flexible fair use exception, but the extent to which fair use permits preservation and related activities is uncertain and has not yet been addressed by US courts.

Acquisition of Digital Content for Preservation

Three of the countries that participated in the study have laws in some form that require the deposit of publicly available copyright materials for the benefit of one or more preservation institutions. Deposit in the Netherlands is by voluntary agreement only. None of the countries, however, has a uniform national system for collection of digital materials, either through a compulsory or a voluntary scheme. While the UK has recently updated its legal deposit legislation, it will require further regulation to extend legal deposit to non-print material. The federal law in Australia only extends to print and US law does not extend to most material made available only online.

The use of technological protection measures to control access to, and use of, digital content may also prevent digital preservation since circumventing such measures may be prohibited by law. Even if preservation institutions are able to circumvent TPMs, the mechanism for doing so may be impractical, as it may involve appealing to the government. The creation and supply of circumvention tools may be illegal anyway. In any case, much digital content is made available through contractual agreements, which over-ride the legal provisions in most circumstances.

Approaches to the Challenges for Digital Preservation

There has been some activity in the study countries related to legal reform. The Gowers Review of intellectual property regimes in the UK made some recommendations on preservation of digital content, which has led to a government consultation on the Gowers recommendations. In the USA, the Section 108 Study Group has made detailed recommendations for changes to US law. There have been a number of reviews of Australian copyright law and policy in recent years. The Netherlands is the exception in that there are neither proposals nor initiatives for legal reform.

Libraries, archives and other preservation institutions have responded in different ways to the challenges that copyright laws currently present for digital preservation. For example, entities in all of the surveyed jurisdictions have embarked upon projects that rely on collaborative agreements between preservation institutions and right holders. These include the National Library of Australia's PANDORA web archiving project, the Dutch Royal Library's deposit agreements with publishers, the British voluntary deposit schemes and US NDIIPP activities. These agreements are important both for the materials they save and for the best practices they engender. Such arrangements are much more prevalent for some types of digital works than for others, however. For example, the most prominent international cooperative archiving and preservation initiatives, LOCKSS and Portico, have largely dealt with scholarly journals so far.

Copyright is a significant legal barrier to the preservation of orphan works in all the study countries. Preserving institutions can make collaborative agreements with rights holders only if they know or are able to find out who the rights holders are and are able to contact them. Many digital works by their nature are prone to becoming orphaned as they are created informally and perhaps collectively. Examples of such works include web pages and wikis. Australia has very limited provision for orphan works. UK law has no general provision for works whose right holder cannot be identified or traced. In the USA, the Copyright Office has made recommendations for reform to deal with orphan works. Orphan works legislation based on these recommendations was introduced in 2006, but not passed by Congress. It has subsequently been re-introduced and is pending.

Conclusions and Recommendations

Digital preservation is vital to ensure that works created and distributed in digital form will continue to be available over time to researchers, scholars and other users. Digital works are often ephemeral, and unless preservation efforts are begun soon after such works are created, they will be lost to future generations. Although copyright and related laws are not the only obstacle to digital preservation activities, there is no question that those laws present significant challenges. Further complicating matters are the evolving commercial markets for digital works, and the apprehension among creators and right holders concerning the impact that further exceptions might have on the market for their works.

Legal reform is needed to ensure comprehensive preservation of the vast range of copyrighted materials now being made available in digital form. The study includes specific recommendations for amendments to the laws in each country. The joint recommendations outlined here focus on amendments to national copyright and legal deposit laws in general that will help to bring these laws into the digital age, whilst being consistent with the legitimate interests of right holders. There are also recommendations for further research concerning issues related to access to preservation copies, and on the relationship of contracts to copyright exceptions (and in particular, to exceptions that facilitate digital preservation).

The study recommended that countries should establish laws and policies to encourage and enable the digital preservation of at risk copyrighted materials. These laws and policies should, at a minimum:

- Apply to all non-profit libraries, archives, museums and other institutions as may be authorised by national law (hereafter, "preservation institutions") that are open to the public, provided they do not undertake these activities for any purpose of commercial advantage.
- Apply equally to all categories of copyright materials, including literary, artistic, musical and dramatic works, as well as to motion pictures and sound recordings.
- Apply equally to copyrighted materials in all media and formats, whether hard copy or electronic, born digital or digitised for preservation.
- Allow preservation institutions to proactively preserve at risk copyright materials before they deteriorate, are damaged or are lost, and before any software or hardware required to access and

use the material becomes obsolete, subject to measures appropriate to protect the legitimate interests of right holders.

- Allow preservation institutions to undertake preservation activities as necessary and in accordance with international best practices for digital preservation, including
 - Reproduction and retention of such copies as may be necessary for effective digital preservation;
 - The serial transfer of copyrighted works into different formats for preservation in response to technological developments and changing standards, and
 - The communication of works within the preservation institution for administrative activities related to preservation, or between the preservation institution and legally authorized third party preservation repositories as necessary for the purpose of maintaining redundant preservation copies to protect against catastrophic loss.

All of the foregoing should be subject to measures appropriate to protect the legitimate interests of right holders.

- Enable relevant preservation institutions comprehensively to preserve copyrighted materials that have been made available to the public in digital form, by means of
 - A legal deposit system,
 - The legal ability to harvest publicly available online content for preservation purposes,
 - Incentives for contractual arrangements for preservation activities, and/or
 - Some combination of the foregoing.

It is also recommended that

- Preservation institutions should work with right holders to develop workable approaches to the digital preservation of copyright materials protected by technological measures such as encryption or copy protection.
- Preservation institutions should develop best practices for digital preservation.

Although not specifically included in the recommendations of the study, participants at the WIPO workshop also identified the issue of orphan works as one that should be addressed with some urgency.

The study recommended that further research should be undertaken on the national level with regard to whether and under what circumstances access to digital preservation copies can be provided without harm to right holders. Finally further research should be

undertaken on the national level to re-examine the interaction between copyright and private agreements as it relates to digital preservation. The research will help in determining whether common approaches to these issues can be developed.

References

Besek, J.M. et al, 2008. *International Study on the Impact of Copyright Law on Digital Preservation*. http://www.digitalpreservation.gov/partners/resources/pubs/wipo_digital_preservation_final_report2008.pdf

Berne Convention for the Protection of Literary and Artistic Works, opened for signature September 9, 1886, 1 B.D.I.E.L. 715. Article 9(2) http://www.wipo.int/treaties/en/ip/berne/trtdocs_wo001.html

Acknowledgements

The authors wish to thank Annemarie Beunen, Erin Driscoll, Michelle Gallinger, Neil Grindley, Helen Hockx-Yu, Scott Kiel-Chisolm, Paul Koerbin, Abigail Potter, Mary Rasenberger, Sarah Waladan.

Developing Preservation Metadata for Use in Grid-Based Preservation Systems

Arwen Hutt*, Brad Westbrook*, Ardys Kozbial*, Robert McDonald**, Don Sutton***

*UCSD Libraries
University of California, San Diego
9500 Gilman Drive, #0175
La Jolla, CA 92037
ahutt@ucsd.edu,
bdwestbrook@ucsd.edu,
akozbial@ucsd.edu

** IUB Libraries
Indiana University Bloomington
1320 East 10th St.
Bloomington, IN 47405
robert@indiana.edu

*** San Diego Super Computer
Center
University of California, San Diego
9500 Gilman Drive, #0505
La Jolla, CA 92037
minor@sdsc.edu, suttond@sdsc.edu

Abstract

Establishing metadata requirements is a key challenge for any attempt to implement a digital preservation repository. The repository's capacity to provide cost-effective, trustworthy services largely derives from the metadata it uses. This paper describes the metadata posited to support services the Chronopolis preservation system will offer at the conclusion of its first year of development.

Chronopolis Overview

The Chronopolis Digital Preservation Framework [1] is a collaborative partnership between the San Diego Supercomputer Center (SDSC), the University of California, San Diego, Libraries; (UCSDL), The National Center for Atmospheric Research (NCAR), and The University of Maryland Institute for Advanced Computer Studies (UMIACS) to establish a digital preservation system within a grid-based network. During the 2008-09 fiscal year, The Library of Congress' National Digital Information Infrastructure and Preservation Program (NDIIPP)[2] awarded funding to the Chronopolis Consortium to build a demonstration preservation data grid containing up to 50 terabytes of heterogeneous data at each Chronopolis node (SDSC, NCAR, UMIACS). The long term goal is to develop a trustworthy digital preservation system offering a spectrum of reliable services to data producers. Short term goals for the first and current development phase include:

- build system infrastructure at three sites (physical machines, software installation, security, software configuration)
- transfer data from depositors
- replicate acquired data across three sites
- develop preservation services utilizing advantages of grid-based networks
- define metadata required to satisfy services

Services

In this first phase Chronopolis project staff is developing basic archiving services, chief of which are:

1. provide replication of files in multiple and geographically dispersed locations
2. provide regular monitoring to identify non-authentic files
3. develop mechanisms for replacing non-authentic files
4. deliver files back to the depositor on request

During this current phase, the team will not implement any of the following services:

1. allow modification of files on our servers
2. provide end user access
3. validate and / or migrate file formats

From a depositor perspective, Chronopolis will provide a data archive that will protect against data loss due to bit decay, system malfunction, natural disaster and vandalism. This will be accomplished by using replication and redundant storage techniques in a grid environment.

Data providers

Data providers for the Chronopolis project include the California Digital Library (CDL), the Inter-university Consortium for Political and Social Research (ICPSR) at the University of Michigan, North Carolina State University (NCSSU) and the Scripps Institution of Oceanography (SIO) at UCSD. All of the data providers are also NDIIPP Partners and the data being ingested into Chronopolis are related to other NDIIPP projects.

CDL, a department of the University of California's Office of the President (UCOP), provides centralized support for digital initiatives that serve all of the libraries in the University of California system. CDL contributed 6 terabytes of data to Chronopolis from its Web-at-Risk project, which has been composed of web crawls of political and governmental web sites over the course of five years. The web crawler packages the data into files of uniform size.

ICPSR is submitting its whole collection of data, consisting of approximately 12 terabytes of data. This

collection includes 40 years of social science research data comprised of millions of small files.

NCSU's data in Chronopolis include approximately 5 terabytes of state and local geospatial data that were collected under the auspices of the North Carolina Geospatial Data Archiving Project, one of the initial eight NDIIPP projects. NCSU is also part of NDIIPP's new multistate effort, which is keenly interested in exchange of digital content among states.

SIO's approximately 2 terabytes of data are made up of data gathered from approximately 1,000 SIO research expeditions during the past 50 years. SIO was able to combine these data into one place with the help of a Digital Archiving (DigArch) research grant from NDIIPP.

The cumulative amount of digital content transferred to Chronopolis' custody is approximately 25 terabytes. These data present themselves in a wide variety of file formats, and the content includes web crawls, geospatial data, social science data and atmospheric/oceanographic data. The Chronopolis team purposely solicited a diverse set of data content and types in order to develop and test Chronopolis' capacity to manage it efficiently and reliably.

Metadata Working Group

The metadata working group was charged with developing metadata specifications for the first phase of Chronopolis development. These metadata specifications have several requirements, they must:

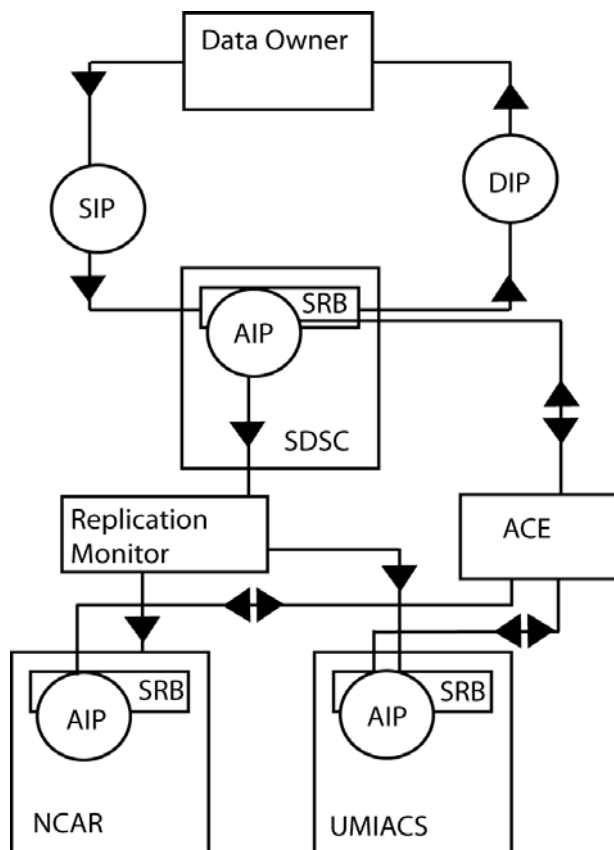
1. Support the services Chronopolis implements in its first phase.
2. Be conformant with community metadata standards.
3. Be extensible to support future development of services.
4. Promote trust between the customer and Chronopolis.

Metadata requirements have been established by working back from services to the events that trigger the services, which is discussed more fully in the ensuing sections.

Workflow Events & Associated Metadata

While it is anticipated that more services will be added in the future, the workflow currently in place within Chronopolis is, in broad strokes, one which the project team expects to follow going forward. The essential stages to the present system are ingest, replication, asset management, and asset retrieval (i.e., delivery back to the customer). These represent broad areas of an object's life cycle, as well as the rudimentary stages of the Reference Model for an Open Archival Information System (OAIS)[3]. A representation of the current system work flow is provided in Figure 1.

Figure 1:



Ingest

Pre-Ingest

Within the Chronopolis project, pre-ingest has required determining the materials to be deposited, agreeing on the format(s) in which they will be submitted, and establishing secure and efficient transfer mechanisms. In addition, part of the pre-ingest process entails configuring the Storage Resource Broker (SRB)[4], the data grid management system used within Chronopolis, to host the submitted content. This involves establishing a collection or hierarchy of collections associated with the depositor.

A characteristic of the current pre-ingest process is that, beyond a few very core pieces of information, there is no need for submitted data to be compliant to a standardized Submission Information Package (SIP)[3] stipulated by Chronopolis. This imposes a non-trivial burden on the Chronopolis system, as there is less control over the form of the submission, as well as on the presence of necessary metadata. The absence of a standardized SIP results in the need to normalize or, where necessary, create the core metadata to enable Chronopolis management services. Without such actions, the Chronopolis system would need to define processes to manage each submitted collection individually, a position that is obviously not scalable or sustainable. That said, the absence of a standardized SIP reduces what might otherwise be a significant barrier for many potential customers. Certainly, its absence enabled the

import of a significant quantity of diverse data within a relatively short time frame.

Metadata

- Depositor name
- Collection name
- Collection structure

Transfer

The process of transferring data is an important component of the overall workflow. The size of the submission(s), as well as whether it is composed of a few very large files or a great number of small files, affects the methods of transfer. Since, as described above, the collections being deposited are extremely diverse, careful attention must be given to the different storage locations and their specifications.

In addition, the submissions have varied in not only their transfer form, but whether the objects are deposited via a push method by the depositor, a pull method by the repository, or a combination of the two. The BagIt standard [5] recently developed by CDL and the Library of Congress, to "simplify large scale data transfers between cultural institutions" [6] is the submission format used for deposit by both CDL and NCSU, and it accounts for 17 of the 25 terabytes deposited in Chronopolis. The BagIt standard is a simple format for transferring digital content focused on the core necessities for efficient and verifiable data transfer. As such, it allows packaging of digital objects with a small amount of accompanying metadata. The core of this metadata is an inventory of content files and a checksum value for each file. It is also possible to point to content files via URLs instead of packaging them 'within' the bag. This configuration is referred to as a 'holey bag' and is an example of a deposit which consists both of pushed content (files within the bag) and pulled content (files which are retrieved via URLs).

Metadata

- File location (when files are transferred via a pull mechanism)

Verification

Regardless of the method by which content is transferred, all files are placed within the staging SRB instance and are subject to an initial audit to assess how complete the transfer was and if all files transferred without corruption. This is done by comparing the transferred files to the manifest to verify that all files were received, and by calculating the checksum value for the file and comparing it to the checksum value calculated before transfer. These quality control procedures allow the identification of any corrupted transfers or missing files. The data provider can then be notified and the appropriate action(s) can be taken.

Metadata

- Original file identifier
- Number of files in the collection
- Size of file

- Checksum algorithm
- Checksum for file

Registration

Once this quality assurance has been accomplished, files are registered within the receiving SRB instance. In most cases metadata is stored within the MCAT, the database system for managing the SRB, but it is not stored as a first class object, like the primary content files themselves. The deposited file's associated MCAT record is supplemented with system level data required for the management of that object, resulting in the creation of the Archival Information Package (AIP)[3], the object to be managed over time.

Metadata

- File identifier
- Date of deposit
- User who uploaded the file(s)
- User's associated group
- Size of file
- Checksum algorithm
- Checksum for file
- Resource where file is stored (information needed so SRB knows how to talk to the resource)
 - Type of resource (e.g., disc, tape)
 - OS resource is running
 - IP address of resource

Archival Storage

There are a number of threats posed to the long term preservation of digital objects. It is possible for problems to be introduced during a processing event, such as transfer to a repository, migration to new media or even delivery back to the data depositor. Failures of media or hardware can cause data loss. Natural disasters can cause catastrophic data loss for an entire repository. And either through error or malicious attack, humans can threaten the integrity of digital objects. There are two important components of protecting digital objects from all of these threats--replication and auditing.

Replication

The Chronopolis Network supports two levels of replication; replication between nodes of the network, also called mirroring, and replication within each node. At present, mirroring between the Network partners provides copies of archived data in three dispersed geographic regions within the United States (the West Coast, East Coast and Rocky Mountains). This level of replication provides protection against data loss through natural disaster, large scale media or hardware failure and human error or attack. Mirroring occurs after ingest is complete, when the AIP is replicated at the other nodes within the network. This process then requires an additional round of quality assurance auditing to insure that all files are present and uncorrupted, and modification of some system level metadata to reflect the content's presence at the replicated node.

In addition, each node can create local replicas of the content managed within the SRB infrastructure. This local redundancy could provide a more efficient protection against data loss due to communication errors in transfer to new media and / or limited media or hardware failure.

This process is facilitated in part by the Replication Monitor[7], a tool developed at the University of Maryland. The tool automatically synchronizes collections between master and mirror sites and logs any actions or anomalies. The Replication Monitor is a tool built on top of the SRB and is a simple web application that watches designated SRB directories and ensures that copies exist at designated mirrors. The monitor stores enough information to know if files have been removed from the master site and when the last time a file was seen. In addition any action that the application takes on files is logged.

Metadata

Data which will match that of the SRB/MCAT from which the data is being replicated from

- Size of file
- Checksum for file
- Checksum algorithm
- Number of files in the collection

Data which will be unique within each node

- Resource where file is stored (information needed so SRB knows how to talk to the resource)
 - Type of resource (e.g., disc, tape)
 - OS resource is running
 - IP address of resource

Data related to replicas

- Date of replication
- File replicated from (node and resource location)
- File replicated to (node and resource location)

Auditing

The second component of archival storage is regular and ongoing monitoring of the files to identify any errors or failures. Regular, scheduled audits are necessary as depositor access to files is infrequent within an archive of this type and so cannot be relied upon for uncovering problems. Auditing allows the identification of data loss in a timely manner so action can be taken to repair or replace the damaged object.

Within Chronopolis this is being done using the Auditing Control Environment (ACE), also developed at the UMD. ACE is a policy driven environment for verifying the integrity of an archives' holdings. ACE provides a two-tiered approach to integrity management. The first tier includes Integrity Tokens and Cryptographic Summary Information (CSI), and the second tier Witness values (See [8][9] for more information about ACE). An important characteristic of ACE is that it is run

independently of the archive, which reduces the chance that a malicious file modification can go undetected since verification information will need to be changed in two independent, and independently administered, systems.

A file must first be registered with ACE. On this registration a token is created which documents integrity information for the file. This, in concert with the CSI and Witness values, is used to conduct regular evaluations of a file, and an archive's, integrity.

Metadata

- Checksum for file
- Version number
- Checksum algorithm
- Last integrity token
- Time stamp
- Aggregation proof
- Last summary information

Dissemination

Within the current project it is expected that Chronopolis will be able to deliver materials back to the depositor in the same form as they were initially submitted. Additionally, Preservation Description Information (PDI)[10] will be provided to document the authenticity of the files. These deliverables will constitute the content of the Dissemination Information Package (DIP)[3].

Metadata

For file submitted

- Size of file
- Checksum algorithm
- Checksum for file

For file returned

- Size of file
- Checksum algorithm
- Checksum for file
- Audit trail documenting events in file's history
 - Deposit
 - Replication
 - Verification
 - Recovery (with a replica when a verification fails)
 - Dissemination

Metadata Packages

Work is now progressing on development of metadata specifications for the AIP and two DIPs. These are focused on documentation of metadata to be collected, created and retained.

As described above, the AIP is composed of metadata elements contributed by the depositor and created by SRB, ACE or the Replication Monitor. These elements are primarily stored within the MCAT database, but also depend upon data within ACE, and so the AIP is not truly a single 'package' in a physical sense, but a logical

one. The system dependencies and distributed nature of the AIP data, necessitates that reference is made to the internal metadata elements for the relevant systems, not that encoding of AIP metadata elements according to an external standards, such as the PREservation Metadata: Implementation Strategies (PREMIS) standard [11], is needed. But while *encoding* according to an external standard is not appropriate, indicating how the AIP metadata meets the requirements established by the community is. Within this context PREMIS is important for its detailed treatment of the metadata elements needed for preservation management, and its grounding within the OAIS framework.

In contrast, it is expected that the Preservation Description Information portion of the DIP will be expressed according to the PREMIS data dictionary and schemas. This package will contain much of the same data elements which make up the AIP, although there will be some variance between the data in the two packages. The DIP must thoroughly document the provenance of the digital object from its ingest into the repository to its dissemination to the depositor.

A mapping of Dissemination Information Package metadata for a file to PREMIS is presented in the chart in Figure 2. It should be noted that this includes the primary metadata which supports Chronopolis services as outlined thus far; it is not intended to be exhaustive of all elements which would be present in a DIP.

Figure 2:

DIP Metadata	PREMIS Elements
Object Entities	
Collection name	linkingIntellectualEntityIdentifier
Original file ID	originalName
Size of file	size
Checksum (pre-ingest)	messageDigestAlgorithm messageDigest messageDigestOriginator=Depositor
Checksum (post-ingest)	messageDigestAlgorithm messageDigest messageDigestOriginator=Repository
Cryptographic Summary	messageDigestAlgorithm messageDigest messageDigestOriginator=Audit control software
File identifier	objectIdentifier
Resource Type	storageMedium
Resource IP	contentLocation
Replica of file	relationshipType=replication relationshipSubType=is equal relatedObjectIdentification
Agent Entities	
Depositor	agentIdentifier agentName agentType=organization
Repository	agentIdentifier agentName agentType=organization

Networked repository	agentIdentifier agentName agentType=organization
Replication monitor software	agentIdentifier agentName agentType=software
Audit control software	agentIdentifier agentName agentType=software
Initiator of file recovery	agentIdentifier agentName agentType=person
Event Entities	
Deposit	eventType=ingestion eventDateTime eventOutcomeInformation
Replication	eventType=replication eventDateTime eventOutcomeInformation
Verification	eventType=fixity check eventDateTime eventOutcomeInformation
Recovery	eventType=replacement eventDateTime eventOutcomeInformation
Dissemination	eventType=dissemination eventDateTime eventOutcomeInformation

Development of the DIP specifications will build on work done during a previous NDIIPP project, Data Center for Library of Congress Digital Holdings: A Pilot Project, a one-year demonstration project to test the feasibility of engaging external partners as service providers to fill digital management needs. During this project, a prototype DIP for transferring preservation responsibility for an object was developed. Chronopolis will expand on that work by modeling an encoding for a more complete audit trail, including representation of mirrored sites, exploring other package formats, and updating the mapping to comply with the recently released PREMIS 2.0 [12].

Conclusion

Implementing a federated digital preservation repository network has required us to closely examine the services to be supported and what metadata is needed to enable them. It is expected that this first phase of development will provide a strong technological, policy and trust foundation upon which Chronopolis can build.

References

- [1] Chronopolis Digital Preservation Framework.
<http://chronopolis.sdsc.edu/>
- [2] The Library of Congress' National Digital Information Infrastructure and Preservation Program.
<http://www.digitalpreservation.gov/>
- [3] Consultative Committee for Space Data Systems. 2002. Reference Model for an Open Archival Information System (OAIS).
<http://public.ccsds.org/publications/archive/650x0b1.pdf>
- [4] Storage Resource Broker.
http://www.sdsc.edu/srb/index.php/Main_Page
- [5] Kunze, J.; Littman, J. and Madden, L. 2008. The BagIt File Package Format (V0.95)
<http://www.cdlib.org/inside/diglib/bagit/bagitspec.html>
- [6] The Library of Congress. 2008. Library Develops Format for Transferring Digital Content, News and Events.
http://www.digitalpreservation.gov/news/2008/20080602_news_article_bagit.html
- [7] SRB Replication Monitor V2.0.
<http://narawiki.umiacs.umd.edu/twiki/bin/view/Main/SrbRepMon2>
- [8] ACE: Audit Control Environment.
<http://narawiki.umiacs.umd.edu/twiki/bin/view/Main/ACEOverview>
- [9] Song, S. and JaJa, J. 2007. ACE: a Novel Software Platform to Ensure the Long Term Integrity of Digital Archives, Proceedings of the Archiving 2007 Conference, May 2007, Washington, DC.
<http://adaptwiki.umiacs.umd.edu/twiki/pub/Lab/Papers/rad71E67.pdf>
- [10] Caplan, P. 2006. DCC Digital Curation Manual: Installment on Preservation Metadata,
<http://www.dcc.ac.uk/resource/curation-manual/chapters/preservation-metadata/preservation-metadata.pdf>
- [11] PREMIS Working Group. 2008. PREMIS Data Dictionary for Preservation Metadata version 2.0.
<http://www.loc.gov/premis/v2/premis-2-0.pdf>
- [12] Lavoie, B. 2008. PREMIS with a Fresh Coat of Paint: Highlights from the Revision of the PREMIS Data Dictionary for Preservation Metadata, D-Lib Magazine, Vol.14 No.5/6.
<http://www.dlib.org/dlib/may08/05contents.html>

Using METS, PREMIS and MODS for Archiving eJournals

Angela Dappert, Markus Enders

The British Library
Boston Spa, Wetherby, West Yorkshire LS23 7BQ, UK
St Pancras, 96 Euston Road, London NW1 2DB, UK
angela.dappert@bl.uk, markus.enders@bl.uk

Abstract

As institutions turn towards developing archival digital repositories, many decisions on the use of metadata have to be made. In addition to deciding on the more traditional descriptive and administrative metadata, particular care needs to be given to the choice of structural and preservation metadata, as well as to integrating the various metadata components. This paper reports on the use of METS structural, PREMIS preservation and MODS descriptive metadata for the British Library's eJournal system.

Introduction

At the British Library, a system for ingest, storage, and preservation of digital content is being developed under the Digital Library System Programme, with eJournals as the first content stream. This was the driver for developing a common format for the eJournal Archival Information Package (AIP) as defined in OAIS [CCSDS 2002].

In order to understand metadata needs, it is helpful to understand the business processes and data structures. eJournals present a difficult domain for two reasons. The first is that eJournals are structurally complex. For each journal title, new issues are released in intervals. They may contain varying numbers of articles and other publishing matter. Articles are submitted in a variety of formats, which might vary from article to article within a single issue.

The second reason is that, the production of eJournals is outside the control of the digital repository and is done without the benefit of standards for the structure of submission packages, file formats, metadata formats and vocabulary, publishing schedules, errata, etc.. As a consequence, systems that handle eJournals need to accommodate a great variety of processes and formats. This paper presents a solution that can accommodate the complexity and variety found in eJournals.

Fortunately, there has been a substantial amount of work over recent years to define metadata specifications that can support complex cases such as eJournals. The Metadata Encoding and Transmission Specification (METS) provides a robust and flexible way to define digital objects ([METS 2006]). The Metadata Object Description Scheme (MODS) provides ways to describe objects, and builds on the library community's MARC tradition ([MODS 2006]). Finally, the Preservation Metadata Implementation Strategy (PREMIS) data dictionary ([PREMIS 2005]) provides ways of describing objects and processes that are essential for digital preservation. These three metadata specifications are all built on an

XML ([XML 2006]) foundation. Their user communities and underlying approaches also have much in common. All of them are content-type independent, which makes it possible to define shared usage guidelines for the various content-types held in the archival store.

Unfortunately, there are many ways to combine these three specifications to provide a complete solution to the problem of defining an eJournal Archival Information Package. This paper explains one approach.

The eJournal Ingest Workflow

Ingesting eJournals requires a complex workflow that needs to be adjusted for each individual information provider's submission process and formats.

Each submission may contain several submission information packets (SIP) as defined in OAIS [CCSDS 2002]. Most SIPs are tarred or zipped files that need to be unpacked and virus checked before they can be processed further.

An unpacked SIP will typically contain content files, descriptive metadata for articles, issues and journals, and manifests listing the content of the SIP with size and hashing information.

Since a SIP may contain one or several issues and articles for one or several journals, each structured according to the information provider's conventions, and possibly containing special issues or supplements, the content needs to be split up into identified packages with well-defined structural relationships. The publisher supplied structural relationships between article, issues and journal objects may have been captured in the directory structure, through file naming conventions or through explicit metadata. In the latter case, issue and journal metadata may have been kept with each article's metadata, or contained as distinct metadata sets that are linked to each other. These relationships are extracted and represented in a uniform way, as specified in the British Library's METS, PREMIS and MODS application profiles.

Publisher supplied metadata may have been represented using in-house formats, standards, or modified standards. We extract metadata either from the publisher supplied metadata or directly from the content. The latter is typically the case for technical metadata. The extracted metadata is then normalized according to the British Library's METS, PREMIS and MODS application profiles. Typically, information providers submit several manifestations of each article. A manifestation is a collection of all files that are needed to create one rendition of an arti-

cle. An HTML manifestation, for example, might consist of the HTML file and several accompanying image, video and sound files. Often the submitted content contains a marked-up representation of an article that, again, may be based on proprietary, standard, or modified standard XML schemas or DTDs.

The result of the ingest and normalization processes is one or more Archival Information Packages (AIPs) that can be stored. Structural relationships, metadata and, possibly, content are normalized in order to ensure uniform search across all digital objects and to guarantee the sustainability of formats, data and structural relationships of the AIPs.

The structure of the AIPs is tied to the technical infrastructure of the preservation system.

Technical Infrastructure

The British Library's technical infrastructure to preserve digital material consists of an ingest system, a metadata management component that may vary for different content-types, and an archival store that is shared for all content-types. They are linked with the existing integrated library system (ILS). This architecture is designed to enable access to resources, as well as to support long-term preservation activities, such as format migrations.

The eJournal *ingest system* under development is highly customizable and can be adjusted for different ingest processes, metadata formats, and ways of bundling and structuring the submitted content files. It extracts and normalizes relevant metadata and content.

The *metadata management component* (MMC) manages all types of metadata in a system-specific form, stores it in a database, and provides an interface for resource discovery and delivery. Since the ILS is designed to hold information on the journal-title and issue levels only, it is necessary to keep all article related information in the metadata management component; the system synchronizes changes to journal and issue information with the ILS.

The *archival store* is the long-term storage component that supports preservation activities. All content files are stored there. All archival metadata (that which goes beyond day-to-day administration) is linked to the content and also placed into the archival store. Even though the metadata in the archival store is not intended to be used for operational access, we consider it good archival practice to hold content and metadata within the same system to ensure that the archival store is complete within itself. This archival metadata is represented as a hierarchy of METS files with PREMIS and MODS components that reference all content files (images, full text files, etc.). The bundle of METS and content files comprises the Archival Information Package (AIP).

METS provides a flexible framework for modeling different document types and scenarios. The example of eJournal preservation will show how complex documents and their relationships are modeled in METS and stored in the system.

AIP Granularity

To understand the design of the system, it is fundamental to know that the objects in the underlying digital store

are write-once in order to support archival authenticity and track the objects provenance; an in-situ update of AIPs in the digital store is not possible. Updated AIPs need to be added to the store and generations need to be managed. (A *generation* corresponds to an update to an object. Words such as *version* or *edition* are heavily over-loaded in the library community.) Updates happen for several reasons, and possibly frequently. A first possible reason is the migration of content files due to obsolete file formats. Second, errors might occur during the ingest process that will result in damaged or incomplete data. Even if effective quality assurance arrangements are made, chances are still high that potential problems are occasionally detected after data has been ingested. Third, updates to descriptive or administrative metadata that is held in the archival store may be needed. Metadata updates might happen in small (e.g., a correction of a typo) or larger scale. Fourth, even though the AIPs are designed to have no dependencies on external identifiers, it is conceivable that updates of other information systems (e.g., the ILS) might affect information stored within the AIP.

In order to deal with updates efficiently, we

- separate structural information about the relationship of the files in a manifestation from the descriptive information and from submission provenance information.
- split logically separate metadata subsets that are expected to be updated independently (journal, issue, article) into separate AIPs.

The eJournal data model, therefore, contains five separate metadata AIPs representing different kinds of objects: journals, issues, articles, manifestations, and submissions. Each one is realized as a separate METS file.

The first three are purely logical objects intended to hold relevant, mostly descriptive, metadata at that level. The remaining two are different.

A *manifestation object* is a collection of all files that are needed to create one rendition of an article. It must not be mistaken for FRBR's definition of a manifestation [IFLA 1998], but is roughly equivalent to the PREMIS representation concept. A manifestation may be original or derivative, such as presentation copies or normalized preservation copies of the article. The manifestation object holds structural information about how its files relate and provenance information about the files' origin.

A *submission object* describes one submission event, including all the tarred and zipped SIP files and a record of all activities performed during ingest. Since data can be lost or corrupted in the ingest process, or a need might arise to ingest the same datasets into a different system, we store the original data as it was provided by the publisher in the archival store linked to from its submission. In the environment of a write-once store, this granularity allows us to update data independently without creating redundant records or content files.

The set of these objects represents a hierarchical data model with well defined links from underlying entities to the direct parent. We store relationships between those AIPs in the AIPs themselves in addition to the metadata management component's database. This ensures that the archival store is a closed system that is consistent and complete within itself.

METS/MODS/PREMIS

Every AIP contains at least one XML file that uses the METS schema. METS provides a flexible framework for modeling different document types and scenarios. Using additional metadata schemas, so called extension schemas, METS can embed descriptive metadata records as well as digital provenance, rights and technical metadata. Figure 1 shows the basic sections of a METS file, which are in use in our system. We store descriptive metadata as a MODS extension to the <mets:dmdSec> section. Provenance and technical metadata are captured as PREMIS extensions to the <mets:amdSec> <mets:digiprovMD> and the <mets:amdSec> <mets:techMD> sections. If the METS file describes content files, then they are identified in the <mets:structMap> section.

Within each AIP there is only one single METS file.

```
<mets:mets TYPE="issue">
  <!-- section for descriptive metadata -->
  <mets:dmdSec>
    <mets:mdWrap MDTYPE="MODS">
      <mets:xmlData> ... </mets:xmlData>
    </mets:mdWrap>
  </mets:dmdSec>

  <!-- section for administrative metadata -->
  <mets:amdSec>

    <!-- section for technical metadata -->
    <mets:techMD>
      <mets:mdWrap MDTYPE="PREMIS">
        <mets:xmlData> ... </mets:xmlData>
      </mets:mdWrap>
    </mets:techMD>

    <!-- section for digital provenance metadata -->
    <mets:digiprovMD>
      <mets:mdWrap MDTYPE="PREMIS">
        <mets:xmlData> ... </mets:xmlData>
      </mets:mdWrap>
    </mets:digiprovMD>

    <!-- section for rights metadata -->
    <mets:rightsMD>
      <mets:mdWrap MDTYPE="MODS">
        <mets:xmlData> ... </mets:xmlData>
      </mets:mdWrap>
    </mets:rightsMD>

  </mets:amdSec>

  <!-- section describing structural relationships -->
  <mets:structMap>
    <mets:div TYPE="issue"
      DMDID="ex01MODS01"/>
  </mets:structMap>
</mets:mets>
```

Figure 1: Embedding MODS and PREMIS in the METS container

METS / MODS / PREMIS Based Data Model

The diagrams in Figures 3, 4, and 5 describe the choices we made for representing the objects, their metadata and their relationships to each other within METS, PREMIS and MODS. As illustrated in Figure 2, METS files are represented as shaded boxes, content files are repre-

sented in white boxes. Relationships that are expressed through

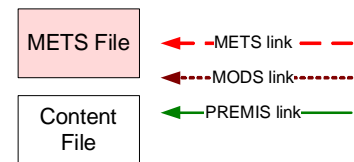


Figure 2: Legend for Figures 3, 4, and 5

METS tags

are shown as dashed arrows; those expressed through PREMIS tags within METS are shown as solid arrows; those expressed through MODS relationship tags within METS are shown as dotted arrows.

Each of the five objects mentioned above is described in a separate METS file that forms a separate AIP. It stores all information about the kind of object and the relationships to other related objects.

Structural Entities: Journal, Issue and Article

Each structural entity, journal, issue and article, is stored in a separate METS file. See Figure 3 for a graphical representation.

Descriptive metadata is expressed using the MODS metadata extension schema to METS and is embedded in a single <mets:dmdSec> section of the METS file. A separate British Library MODS profile describes the elements in use and their meaning.

We use the MODS “host” link to express the hierarchical parent/child relation between journal, article, and issue.

The link uses a unique identifier that is stored within the parent object using the <mods:identifier> element. In our implementation, the identifier is called an MMC-ID (Metadata Management Component identifier).

```
<mods:mods>
  <mods:identifier type="MMC-ID">
    Identifier_of_object
  </mods:identifier>
  <mods:relatedItem type="host">
    <mods:identifier type="MMC-ID">
      Identifier_of_parent
    </mods:identifier>
  </mods:relatedItem>
</mods:mods>
```

Links from child to parent are suitable for systems with a write-once approach. The child objects are updated with greater frequency; for example, each new issue links to the journal. If the link was represented in the other direction, it would be necessary to create a new generation of the journal object for each issue. There are two issues that must be considered when implementing this approach. First, the identifier for the parent must be available before the child’s AIP can be defined and ingested. Second, additional indices must be used to support efficient traversal and retrieval.

As the issue or journal AIPs do not contain information about the order of the articles, and as articles may be ingested out of sequence, the position of the article within an issue must be stored within the article’s descriptive metadata. The <mods:part> element contains machine and human readable sorting information.

```
<mods:part order="w3cdtf">
  <mods:date encoding="w3cdtf">1984
</mods:date>
```

```

<mods:detail type="volume">
  <mods:number>38</mods:number>
</mods:detail>
<mods:detail type="issue" order="1984038030">
  <mods:number>3</mods:number>
</mods:detail>
</mods:part>

```

This information can be used to create a table of contents.

The MODS `<mods:relatedItem>` element is also used to express a range of descriptive relationships between different objects. For example, the relationship among articles that comprise a series is expressed using the value “series” for its *type* attribute; the relationship between a journal published under a new name and the journal as it was previously known is expressed using the value “preceding” for its *type* attribute. Some of these relationships may refer to objects that are held outside the archival store using suitable persistent identifying information.

```

<mets:amdSec>
  <mets:digiprovMD>
    <mets:mdref
      MDTYPE="OTHER"
      OTHERMDTYPE="Preservation Plan"
      LOCTYPE="OTHER"
      OTHERLOCTYPE="MMC-ID" ... />

```

Provenance Metadata for Structural Entities

Long-term preservation requires us to keep a careful record of events related to digital material. Events might impact the data being preserved; data can be lost, corrupted or modified by an event. Some events won’t impact the data itself, but extract information from the data to be used during its processing. Information about events is stored in the AIP’s digital provenance metadata section using the PREMIS schema. Events can be associated with any object type.

As mentioned earlier, in a write-once environment, updates to metadata require the creation of a new generation of the structural entity. In this case, a new AIP for the journal, issue or article is created. This model results in several AIPs representing the different generations of one logical object. While the MODS section within the METS file defines a unique MMC-ID identifier for the logical journal, issue or article object, we need an identifier that is unique to the specific AIP of the object’s generation. Journal, issue or article AIPs have a single PREMIS section under the

updates to metadata require the creation of a new generation of the structural entity. In this case, a new AIP for the journal, issue or article is created. This model results in several AIPs representing the different generations of one logical object. While the MODS section within the METS file defines a unique MMC-ID identifier for the logical journal, issue or article object, we need an identifier that is unique to the specific AIP of the object’s generation. Journal, issue or article AIPs have a single PREMIS section under the

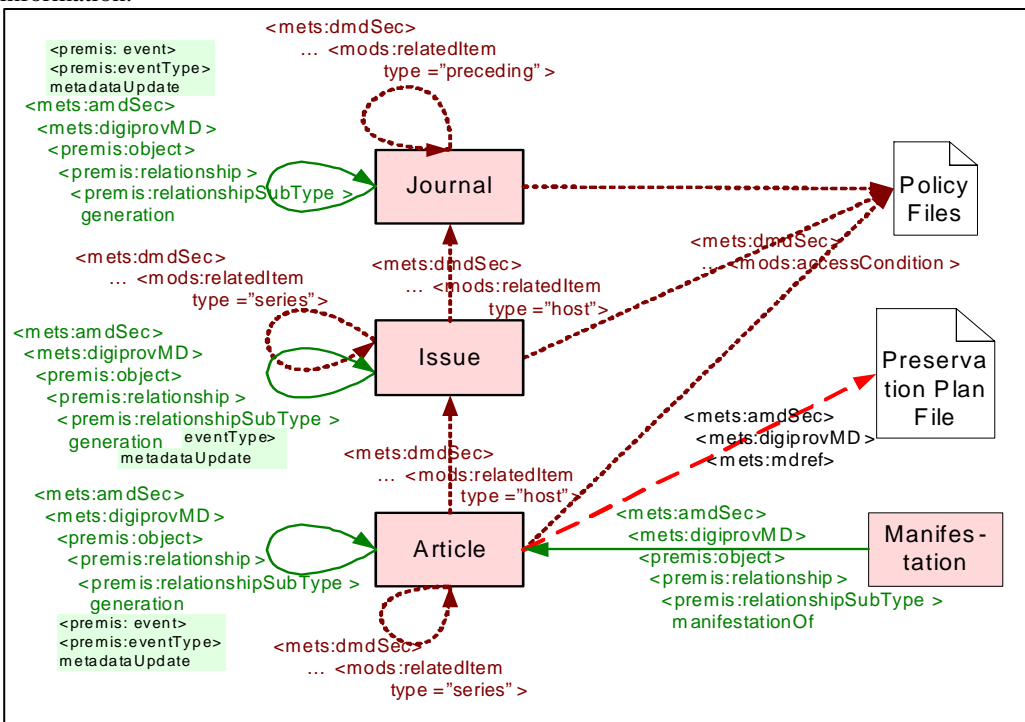


Figure 3: Data model - Structural Entities

For preservation purposes, we also store intrinsic, non-volatile rights information. This includes copyright information as well as license information. License information is stored in a separate policy AIP. Thus, every object that is licensed or acquired under a single policy can refer to the same policy object. This makes it easy to update rights information for several objects at a time without changing a large number of AIPs. The `<mods:accessCondition>` element is used to store the link to the policy file.

```

<mets:dmdSec> ...
<mods:mods>
  <mods:accessCondition
    type="GoverningLicense"
    xlink:href="http://xxxxx"/>

```

Similar considerations apply to Preservation Plans to which article objects link by the `<mets:mdref>` link.

subsection that stores the AIP’s identifier within the `<premis:objectIdentifier>` element as a generation identifier, called MMC-ID+.

```

<mets:amdSec>
  <mets:digiprovMD>
    <premis:object>
      <premis:objectIdentifier>
        <premis:objectIdentifierType>
          MMC-ID +
        </premis:objectIdentifierType>
        <premis:objectIdentifierValue>
          MMC-ID.20070909:3
        </premis:objectIdentifierValue>
      </premis:objectIdentifier> ...

```

The logical object can, hence, be addressed via the MMC-ID stored in MODS; the AIP that represents a

certain generation of the logical object together with information about its digital provenance can be addressed via the MMC-ID+ stored in PREMIS. This is true to our attempt of keeping logical, descriptive information in MODS, and digital provenance information in PREMIS. Each relationship in our data model is expressed via the appropriate identifier type.

The MMC-ID+ identifier is derived from the MMC-ID identifier, concatenated with a colon and a version number; it is unique for the AIP.

A `<premis:relationship>` “generation” link identifies the predecessor’s AIP, and specifies the “metadataUpdate” event in which it had been derived from it and the `<premis:agent>` that executed the event.

```

<mets:amdSec>
  <mets:digiprovMD> ...
  <premis:object> ...
  <premis:relationship> ...
  <premis:relationshipSubType>
    generation ...
  <premis: event>
    <premis:eventType>metadataUpdate ...

```

All digital provenance metadata is captured using the PREMIS extension schema, and is stored within the administrative metadata section of METS. Other types of provenance metadata used will be discussed below.

Manifestation

Each manifestation of an article is stored in a separate METS file. A manifestation links its actual content files together, records all events that have happened to its content files (such as uncompressing, migrating, extracting properties) and links to related versions of those files (such as the original compressed file, or the file that was the source for the migration or normalization). Manifestations are linked to their article and submission objects

using the PREMIS extension schema. See Figure 4 for a graphical representation of these properties.

Each manifestation has one `<mets:amdSec>` that is dedicated to holding information about itself, and one for each of its files. The manifestation’s `<mets:amdSec>` section contains the unique identifier for the manifestation and links to the article and submission objects using their MMC-ID using PREMIS

```

<mets:amdSec>
  <mets:digiprovMD> ...
  <premis:object>
    <!-- identifier of the manifestation -->
    <premis:objectIdentifier>
      <premis:objectIdentifierType>
        MMC-ID+
      </premis:objectIdentifierType>
      <premis:objectIdentifierValue>
        MMC-ID.12345:1
      </premis:objectIdentifierValue>
    </premis:objectIdentifier>
    <premis:relationship>
      <premis:relationshipType>
        Derivation
      </premis:relationshipType>
      <premis:relationshipSubType>
        manifestationOf
      </premis:relationshipSubType>
      <premis:relatedObjectIdentification>
        <premis:relatedObjectIdentifierType>
          MMC-ID
        </premis:relatedObjectIdentifierType>
        <!-- identifier of the article -->
        <premis:relatedObjectIdentifierValue>
          MMC-ID.32596:1
        </premis:relatedObjectIdentifierValue>
      </premis:relatedObjectIdentification>
    </premis:relationship>

```

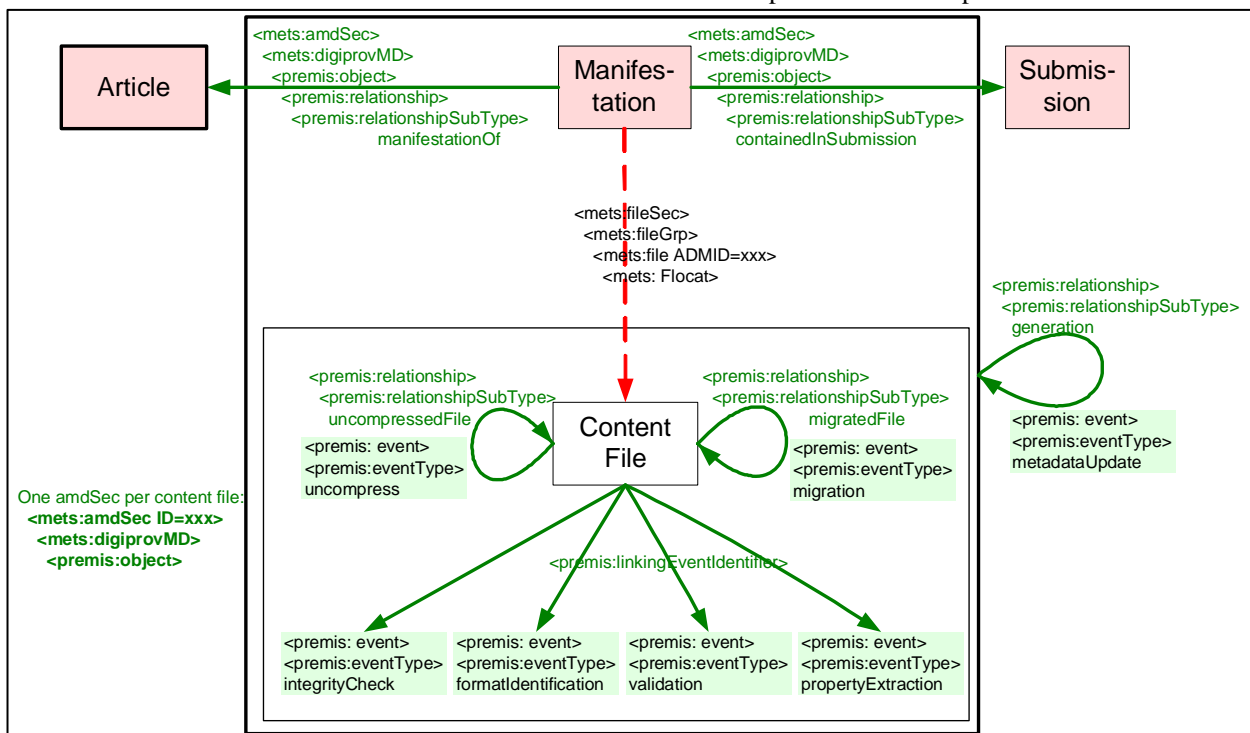


Figure 4: Data Model - Manifestation

Similarly the link to the manifestation's submission object is realized through a <premis:relationship> <premis:relationship-SubType> "containedInSubmission".

The <mets:fileSec> section is used to identify each content file of the manifestation by defining a METS ADMID administrative identifier for it, and to link to the AIPs where these files are actually stored. Each content file receives its own <mets:amdSec> section that is linked to the file by the ADMID that was defined in the <mets:fileSec>. This <mets:amdSec> section stores preservation metadata for content files.

Preservation metadata should support authenticity, understandability and identity of digital objects in a preservation context and represent the important information to preserve digital materials over a long term [PREMIS March 2008]. The METS schema does not have a section dedicated to preservation metadata. Instead it splits preservation metadata into technical, digital provenance, source and rights metadata. Therefore, preservation metadata represented in PREMIS needs to be split up and distributed over these sections. General considerations for this decision process have been discussed in [PREMIS June 2008] and [Guenther 2008].

Fixity and format information for files are regarded as technical information. Therefore the appropriate <premis:object> element containing this information is stored within the <mets:techMD> section. The digital provenance information contains basic identification and provenance information (relationships and events, as well as their attached agents) and is stored within the <mets:digiProvMD> section. As the PREMIS-container element <premis:premis> is not used, this is in accordance with the current METS-PREMIS guidelines [PREMIS June 2008].

As some of the metadata described in the PREMIS data dictionary is mandatory and the XML file won't validate without incorporating this metadata in every PREMIS section, the object-identifier as well as the object category are repeated in each section.

Provenance Metadata for Manifestations

Similar to structural entities, manifestations can have a "generation" <premis:relationship> that identifies a predecessor AIP and a "metadataUpdate" event if correction of metadata has been necessary.

In contrast, a different kind of relationship, however, is not realized as provenance metadata. When the actual semantic content of an AIP gets updated, it is usually regarded as versioning of content. Within the eJournal context a new version is created whenever the publisher decides to publish a corrected or an enhanced version of an article. From the preservation system's perspective this article is seen as a separate expression and a new AIP is created for the article as well as for its manifesta-

tion. The link between two expressions is not regarded as digital provenance metadata. For this reason the link to the previous version is stored in the descriptive MODS metadata record of the new version.

Provenance Metadata for Files

There are two different kinds of events for files: those which are side-effect free and capture information about a file, and those which result in the creation of a derivative file.

Side-effect free events include identification and validation of the file's format, extraction of properties or metadata, and validation of the file's contents to ensure its authenticity when it is disseminated, by checking and comparing the data against stored metrics, such as checksum values. These events are represented in the file's <mets:amdSec><mets:digiprovMD> section in its manifestation's METS file. Storing e.g. the metadata extraction process as an event lets us store the metadata extraction software used during this process as a related agent. As the event does not change the file, no relationship of the file to other objects is stored in the PREMIS metadata.

Derivation events produce a new bytestream while preserving its significant properties and semantic content. This will, for example, happen if an obsolete file format is regarded as "at risk" and a migration has to take place. In this case a new manifestation is created. A "migrated-File" <premis:relationship> links back from each file that results from the migration to each file that fed into the migration (One or several files can be migrated to one or several files). The "migration" event in which it had been derived and the <premis:agent> that executed the migration are recorded with the resulting file.

All files of a manifestation are referenced in its <mets:fileSec>. If a new manifestation is created during a "migration" event, in its <mets:fileSec> it may reference some unchanged files and some that resulted from the migration event. The un-affected files won't have any related file relationships.

A relationship and event similar to the "migration" event is recorded for an "uncompress" event and links from the uncompressed files to the compressed file.

Usually the event outcome should be successful. If a "migration" or an "uncompress" event fails, the file will not be ingested. However, for certain events a negative outcome will not prevent ingest. For example, if the validation of a file cannot be carried out successfully (the validation-event fails), it is handled as an exception and attempts are made to fix the file. If this is not successful, it might be decided to ingest an invalid file just to make sure that the manifestation is complete. Further attempts at fixing the file can possibly be made in the future. This event outcome is recorded within the PREMIS event.

Submission

A submission object describes a single submission event. This includes all the tarred and zipped SIP files as they have been submitted by a publisher, and a record of all activities performed during ingest. See Figure 5 for a graphical representation of these properties. The structure and relationships of the submission object are very similar to the manifestation object. Rather than content files, it references SIPs in its `<mets:fileSec>` and each SIP has a `<mets:amdSec>` of its own to hold its provenance metadata.

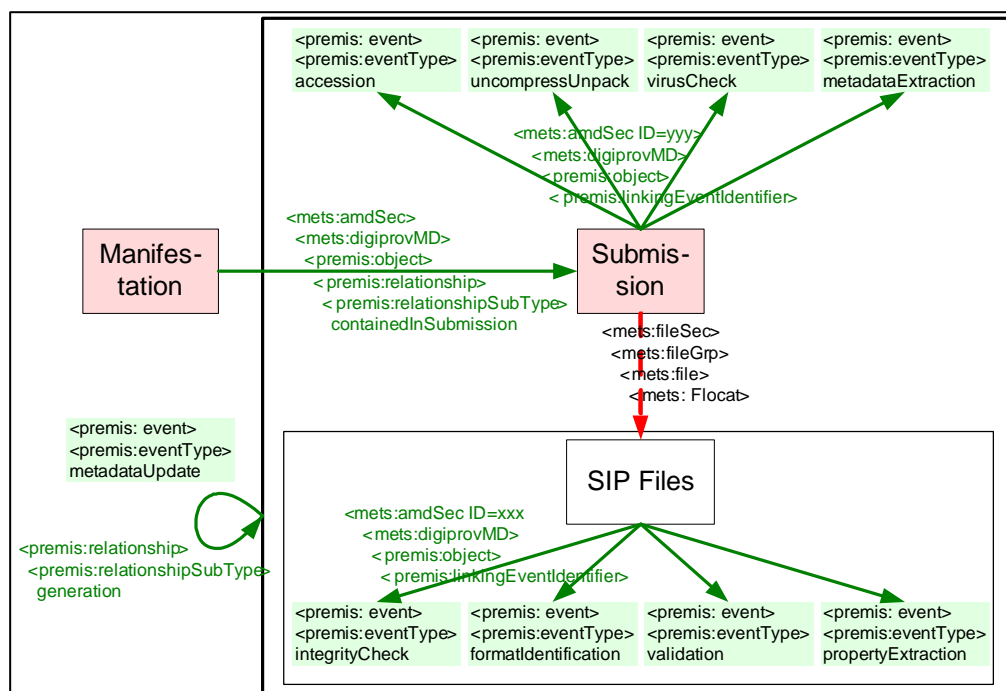


Figure 5: Data Model - Submission

Provenance Metadata for Submissions

Similarly to structural entities and manifestations, a submission object can have a “generation” `<premis:relationship>` with a “metadataUpdate” `<premis:event>`.

All other events recorded for a submission object are free of side-effects. Events such as “accession”, “uncompressUnpack”, “metadataExtraction” and “virusCheck” should theoretically be recorded on a file level, but are actually recorded at the submission level in order to avoid redundancy, since they are identical for all files of a submission.

Provenance Metadata for SIPs

The `<mets:amdSec>` associated with each SIP file has the same side-effect free events as were described for content files in manifestations. It does not contain any events with side-effects or relationships to other files.

METS, PREMIS and MODS Trade-offs

Several metadata elements can be represented in either or several of the metadata schemas. When choosing between them it is helpful to consider that the purpose of the metadata schemas are very different. METS describes a document, while PREMIS stores preservation data for the document or for certain parts (files) of it, and MODS captures descriptive information. Some of the metadata that is captured can be used for several different purposes.

Basic technical metadata, for example, such as checksums and file sizes, are important for preservation purposes but are also part of a complete and detailed description of a digital document. Appropriate elements are available in both schemas (for example,

`<premis:object>``<premis:objectCharacteristic>`
`<premis:size>` and
`<mets:fileSec>``<mets:fileGrp>``<mets:file SIZE=...>`

as well as

`<premis:object>``<premis:objectCharacteristic>`
`<premis:fixity>``<premis:messageDigest>` and
`<mets:fileSec>``<mets:fileGrp>`
`<mets:file CHECKSUM=...>`)

It is envisaged that this information is used in use cases that access either the METS or the PREMIS metadata portions separately. We therefore store this identical information redundantly in METS and PREMIS. Additionally, it was desirable to be able to store several checksums in a repeatable element, such as in PREMIS, rather than in a non-repeatable attribute, such as offered by METS.

For file format information our considerations were as follows. While METS only stores the MIME-type of a file, PREMIS permits referencing an external format registry. For eJournals the PRONOM database is used and referenced. The MIME-type is usually sufficient to disseminate and render a file (e.g., the MIME type needs to be incorporated in the http-header when transferring files). But for preservation purposes further information about the file format, such as the version or used compression algorithm, might be very important. In theory the MIME type could be extracted from the PRONOM registry, but every dissemination would require a request to the PRONOM database. Storing the data redundantly is, therefore, convenient, especially as there was no concern about data becoming inconsistent in our write-once archival store. While the `<mods:physicalDescription>` element offers the possibility of specifying technical properties, we decided to keep all technical metadata together in METS or PREMIS where they would be used together. Using the relevant MODS subelements offered no advantage over the more fit-for-purpose elements in PREMIS and METS. We therefore chose not to use MODS on a manifestation or file level at all.

Even though the relationship between a manifestation object and its article object can be regarded as a hierar-

chical one and could be recorded via a <mods:relatedItem> link, we did not want to introduce a MODS section for manifestations just for holding it. Instead the "manifestationOf" <premis:relationship> element is used within the administrative metadata section. Rights information in our AIPs is not intended to be actionable, in the sense that it does not directly support any repository function, such as access or preservation. Rather it is of an archival, descriptive nature. We, therefore, capture it in MODS rather than PREMIS in order to keep it together with other descriptive information. MODS rights information that is of an administrative nature and might change but is still considered archival, such as embargo information, is stored in the <mets:amdSec><mets:rightsMD> section, whereas descriptive rights information, such as the persistent copyright statement, is kept with other descriptive metadata in the <mets:dmdSec>.

An event that affects several objects is recorded in each affected object's <mets:amdSec>. To create a complete set of metadata, the related agent - the software that executed the event - is stored within the same <mets:amdSec>. Unlike proposed in the current version of the METS-PREMIS guidelines ([PREMIS June 2008]), the <premis:agent> is stored redundantly within each PREMIS section of the same METS file. As each PREMIS section contains a complete set of metadata for a file, extracting, storing or indexing it for preservation purposes becomes very easy.

PREMIS 1.0 versus 2.0

Our current implementation uses version 1.1 of the PREMIS data dictionary ([Premis 2005]) and the corresponding XML schema. After version 2.0 was released in March 2008 ([PREMIS March 2008], [Lavoie 2008]), the impact of changes on the current AIP format were investigated.

Neither the fundamental data model of PREMIS, nor the event and relationship information have changed. The most important change is the possibility of using extensions from within PREMIS that permit embedding of metadata from other metadata schemas. Some elements used in the AIPs could be refined within PREMIS using an additional metadata schema. The event outcome, as well as the creating application, the object characteristics, and the significant properties could be described in more detail.

For us, the <premis:objectCharacteristicsExtension> might beneficially be used to capture further, object or format-specific, technical metadata for a file. Currently this data is stored in a <mets:techMD> technical metadata section using the JHOVE schema. If it is only used for preservation purposes, it might be useful to move it to the <premis:objectCharacteristicsExtension> instead.

Bigger changes have been made in the XML schema. It does not only support the additional elements, but also defines abstract <premis:object> types, and creates special instances for *representation*, *file* and *bitstream*. These instances allow the mapping of the data dictionary's applicability and obligation constraints to the XML

schema and ties them to the object type. This might improve and simplify the validation process

Conclusion

No single existing metadata schema accommodates the representation of descriptive, preservation and structural metadata. This paper shows how we use a combination of METS, PREMIS and MODS to represent eJournal Archival Information Packages in a write-once archival system.

References

- METS 2006. *Metadata Encoding and Transmission Standard (METS) Official Web Site. Version 1.6.* <http://www.loc.gov/standards/mets/>
- PREMIS Working Group, May 2005. *Data Dictionary for Preservation Metadata: Final Report of the Premis Working Group, version 1.0.* <http://www.oclc.org/research/projects/pmwg/premis-final.pdf>
- MODS 2006. *Metadata Object Description Schema. Version 3.2.* <http://www.loc.gov/standards/mods/>
- Bray, T. et al. eds. 2006, *Extensible Markup Language (XML) 1.0 (Fourth Edition).* <http://www.w3.org/TR/2006/REC-xml-20060816/>
- CCSDS, January 2002. *Reference Model for an Open Archival Information System (OAIS). CCSDS 650.0-B-1, Blue Book* (the full ISO standard). <http://public.ccsds.org/publications/archive/650x0b1.pdf>
- IFLA Study Group on the Functional Requirements for Bibliographic Records, 1998. *Functional requirements for bibliographic records : final report.* München: K.G. Saur, 1998. (UBCIM publications; new series, vol. 19). ISBN 3-598-11382-X.
- PREMIS Editorial Committee, March 2008. *PREMIS Data Dictionary for Preservation Metadata, version 2.0.* <http://www.loc.gov/standards/premis/v2/premis-2-0.pdf>
- PREMIS in METS Working Group, June 2008 *Guidelines for using PREMIS with METS for exchange.* Revised June 25, 2008. <http://www.loc.gov/standards/premis/guidelines-premismets.pdf>
- Guenther, R., 2008. *Battle of the Buzzwords; Flexibility vs. Interoperability When Implementing PREMIS in METS.* D-Lib Magazine, July/August 2008, Vol. 14 No. 7/8, doi:10.1045/dlib.magazine, ISSN: 1082-9873. <http://www.dlib.org/dlib/july08/guenther/07guenther.html>
- Lavoie, B. 2008. *PREMIS with a fresh coat of paint: Highlights from the Revision of the PREMIS Data Dictionary for Preservation Metadata.* D-Lib Magazine, May/June 2008, Vo. 14 No. 5/6, ISSN 1082-9873. <http://www.dlib.org/dlib/may08/lavoie/05lavoie.html>

This article was also published in D-Lib Magazine, September/October 2008, Vol. 14, No. 9/10.

Harvester results in a digital preservation system

Tobias Steinke

Deutsche Nationalbibliothek
Adickessallee 1
60322 Frankfurt
Germany
t.steinke@d-nb.de

Abstract

In the last few years libraries from all around the world have build up OAIS compliant archival systems. The information packages in these systems are often based on METS and the contents are mainly e-journals and scientific publications. On the other hand Web archiving is becoming more and more important for libraries. Most of the member institutions of the International Internet Preservation Consortium (IIPC) use the software Heritrix to harvest selected Web pages or complete domains. The results are stored in the container format ARC or the successor WARC. The files' quantity and the sizes of these archival packages are significantly different than those of the other publications in the existing archiving systems. This challenges the way the archival packages are defined and handled in current OAIS compliant systems.

This paper compares existing approaches to use METS and Web harvesting results in archival systems. It describes the advantages and disadvantages of treating Web harvests in the same way as other digital publications in dedicated preservation systems. Containers based on METS are set side by side with WARC and its possibilities.

Background: Preservation systems and Web archiving

In the last few years cultural heritage institutions like national libraries began to build up dedicated archival systems for digital preservation. Coming from the traditional collection of books and journals the focus was on similar digital entities like e-theses, e-journals and digitized books. These items can be in a variety of file formats and quantities but each single object is clearly defined and contains seldom more than a few hundred files. Nearly all of the archival systems are more or less designed according to the OAIS reference model, which identifies components and tasks of such a system. To fulfill the task of preservation it is necessary to ensure access to the content of the objects even when software and hardware will change completely. In the OAIS model the needed activities are called *Preservation Planning*. Current implementations try to do this basically by the strategies migration and emulation. The

basis for both strategies is supporting metadata especially about the technical aspects of each archived object and file.

On the other hand cultural heritage institutions have to face a completely new challenge: The collection and archiving of Web pages. Depending on the institution and existing legal deposits, this could include certain sub domains, pages related to a specific topic or a complete top-level domain like .fr. The common way to collect the pages is to use software called harvester. This automatic program gets an address to start with and then follows every link on each page within given parameters. The result is either saved in separate files according to the original file formats (HTML, JPEG, etc.) or in one aggregated file. One of the most commonly used harvesters is called Heritrix. It saves the results in a aggregated format called WARC. WARC is an ISO draft which contains the files itself and metadata about the harvest activity.

As the process of collecting the Web pages and giving access to them is a challenging process for itself, the actual storage is currently often done without the same requirements for preservation as for other digital objects. Existing archival systems for digital preservation have often not been designed to deal with the complexity of Web pages. Strategies for preservation may be difficult to accomplish on the scale of Web harvester results.

Rebecca Guenther and Leslie Myrick wrote an article in 2006 about the way Web harvester results could be handled as archival packages with the metadata standards METS and MODS [1]. Since then the WARC format became relevant as a more advanced format for Web harvester packages including metadata and on the other hand dedicated archival systems for digital preservation - like the one developed in the German project kopal - became more sophisticated.

Preservation systems and the object model

The ISO standard "Reference Model for an Open Archival Information System (OAIS)" [2] describes an abstract model of an archival system dedicated to long

term preservation. This reference model and especially its functional model define the functional entities and terms commonly used in all developments of digital preservation systems. The objects in the OAIS model are called Submission Information Package (SIP) at the moment of ingest, Archival Information Package (AIP) within the archival storage and Dissemination Information Package (DIP) for the access. Each Information Package is a conceptual container of content information and Preservation Description Information (PDI). The OAIS model does not define or restrict what the content information actually is.

One of the first implementations based on the OAIS reference model was the e-Depot of the National Library of the Netherlands [3]. It was conceived for digital publications, which are mostly PDF files. Therefore the object model was suited to handle single files and low complexity objects.

The German project kopal and the Universal Object Model

The German project “kopal: Co-operative Development of a Long-Term Digital Information Archive” (2004 - 2007) [4] used the same core system (DIAS by IBM) as the e-Depot, but enhanced it with a new object model to enable more complex objects and support the preservation strategy of file format migration. Although the object model was conceived to be able to handle all kinds of file formats and objects with hundreds of files, the focus was still on digital publications by commercial publishers, scientific publications and digitized books.

The object model defined for the kopal project is called Universal Object Model (UOF) [5]. The idea of the UOF is to define an information package, which should contain all files of one logical unit (e.g. a book, a thesis, an article) and all necessary metadata to enable preservation actions like migration. Descriptive metadata could be part of the package, but only the preservation metadata is mandatory. The UOF should also be self-sufficient in a way, that is to be suitable to enable exchange of objects between different archival systems. The package itself is a file container (ZIP or a similar format) and a XML metadata file conforming to the Metadata Encoding & Transmission Standard (METS) [6].

METS is a widely used standard to encode different metadata information and structural information about a digital object in a XML file. It is very generic in a way that there is no restriction on the kind of metadata to be included. Therefore the concept of profiling was established to define restrictions for specific use cases. Rebecca Guenther and Leslie Myrick describe in their article [1] the METS profiles of the project MINERVA which includes descriptive metadata in the format MODS and hierarchal structural information. The METS profile for the UOF demands preservation metadata in the format LMER [7] but allows all kinds of descriptive metadata. All files of the package must be listed in the METS file and there should be a record of technical

information included for every file. Structural information could be of any complexity, but this should be restricted on files within the package.

Web archiving and harvester

Web pages are part of the cultural output of our society and therefore cultural heritage institutions feel the obligation to collect them like any other digital publications. But the structure of the Web is global and there are no clear national borders in the virtual space. The traditional collection policy of national libraries to collect everything from or related to their own country is difficult to apply to Web pages. This problem is addressed by restricting the collection to pages of a certain top-level domain (e.g. .de, .fr, .uk). As an alternative or in addition there could also be a selective approach to collect topic-related.

Another problem is the dynamic character of the Web. There is never a fixed or final state of a Web page. The content of a Web page could be changed at any point in time. The content could also be dynamic itself, computed at the time of access based on input by the user. As a result, collections of Web pages are always time specific snapshots of certain states. It is not possible to collect “The Web”.

The actual collection is done by a harvester (a.k.a. crawler). Starting with a URL these programs follow each link on a page and save every file on their way. The International Internet Preservation Consortium (IIPC) [8] was founded by national libraries and the Internet Archive to collaborate on preserving the Internet content for future generations. Currently it consists of 38 member institutions from all over the world. One of the projects of the IIPC is the (further) development of the open source harvester Heritrix [9]. It uses very sophisticated methods to fetch as much content as possible. The result is stored in ARC files, which are containers for the collected files and the additional information about the harvesting itself.

The WARC format

The WARC format [10] was developed as a successor of the ARC format. It currently exists in a draft status and was submitted as an ISO standard. Every WARC file is a container of records. The records can contain the unchanged binary files of the page (e.g. HTML, JPEG, GIF), general information about the Web crawl, network protocol information, revisitation information (about changes since the last snapshot of the same pages), conversions (migrated file versions) and metadata about each file. The metadata could be WARC specific, Dublin Core or conforming to any other schema. Heritrix will generate one or more WARC files for each crawl depending on a configurable WARC file size.

Approaches to use METS in Web archiving

Most of the institutions which use Heritrix store the resulting ARCs in a file based system and use software like Wayback [11] to give their users access to the stored snapshots. The focus is on managing the harvesting process. Existing preservation systems are separated from these processes.

METS is widely used for SIPs in OAIS compliant archival systems. As the result of a harvester like Heritrix is already a container (ARC or WARC), the containers could be referenced in the METS files or each file in the containers could be referenced individually.

METS in the MINERVA project

The MINERVA project [12] at the Library of Congress (USA) established an archive of event-related collections of Web pages. Although this project was not primary about preservation, Rebecca Guenther and Leslie Myrick [1] described a concept of METS and preservation information for MINERVA. They argue that in order to handle the complexity of the Web material it is necessary to define two METS profiles: One to describe the levels of aggregation and one for every capture. The structural map of the aggregate-level METS files consists of pointers to lower-level METS objects. MODS is used in the METS file to describe the intellectual object on the aggregate-level. The METS files on the capture level includes MODS for page-specific content information, several metadata schemas for technical information on file level and PREMIS for preservation information. The Structural Map and Structural Link section of METS could be used to reflect the links on each HTML page.

METS in the Web Curator Tool project

The Web Curator Tool (WCT) project [13] is a collaborative effort by the National Library of New Zealand and the British Library, initiated by the IIPC. Its purpose is to manage the selective Web harvesting process. A SIP specification [14] was developed for the use case of submitting the results of a harvesting process to an archival system. The SIP contains all ARC files of a crawl, selected log and report files of Heritrix and a METS file. The ARC files and Heritrix files are referenced within the METS file. The Metadata in the METS file conforms to a specific WCT schema and includes information about the crawl, owner data, agency data, descriptive information and permission data. There is no list of the files within the ARC files or technical information about these files in the METS file. The Structural Map is just a plain list of the ARC files and the Heritrix files.

Preservation strategies and Web archiving

An archival system for digital preservation should be focused on ensuring the access to its content for the unpredictable future. Software and hardware will change and no file format will be supported forever. The two common strategies to face this challenge are migration

and emulation. Migration is the conversion of file formats to currently accessible file formats. Emulation is the recreation of another system environment on a currently used system environment. For both strategies it is essential to record as much information as possible about the technical parameters of the archived objects. This is done by generating metadata and storing it together with the content files. METS could be used to build information packages of metadata and content files.

Migration of Web harvester results could be difficult to handle. One crawl can produce thousands of files. A lot of these files are HTML files with links to other files. In case of the migration of one format to another, not only all affected files have to be change but also all HTML files linking to these files. The approaches of the MINERVA project and the UOF enable the recording of technical information for every file and of dependencies between the files in a METS file. In principle this is a good basis for the migration task. But the practical problems of performing all necessary activities (conversions, checks, error corrections) for objects with thousands of files remain. It may also be technically challenging to generate the metadata and the resulting huge METS files on this scale. Migration on the basis of the WCT METS files might be impossible, because there is no information about the technical aspects of the single files within the ARC files. But this approach is helpful for migrations of the ARC files (e.g. to WARC files).

Emulation for Web harvester results could be an easier task than to emulate complete computer systems. Web pages are in principle designed to work on any Web browser of a certain time period. There are dependencies of certain media plug-ins, software specific restrictions and machine related parameters (performance, memory size) but these are harmless compared to the complexity of the emulation of a specific computer configuration. For the emulation approach it is important to know the time period of the crawl and the circumstances of the harvesting process. This is provided in a useful way by the WCT SIP specifications. The ARC files bundle the unchanged content files and the metadata and the reports give the needed information. The MINERVA METS files on the aggregate level would also provide the information. But it could be difficult to hand over all files of one crawl to the emulator. A few ARC files might be easier to handle than thousands of different files. The UOF was not yet used for Web harvester results. If the ARC files were chosen as content files and the technical metadata within the LMER sections described the crawl, the resulting UOF METS files would be similar to the WCT ones.

On the other hand the new WARC format already offers all needed information for the emulation and even a mechanism to store migrated file versions within the container. But WARC files need to be managed in an archival system and therefore a structural wrapper like METS could be helpful. The provided information within the WARC files could be easily extracted to build up METS files which could even support both preservation strategies similar.

Summary

Web archiving is a new challenge for the preservation community. Existing OAIS compliant archival systems use METS and preservation metadata to support preservation strategies like migration and emulation. These concepts could be used for Web archiving as well but a re-design or enhancement of the METS based object models might be necessary. The introduction of the WARC file format offers additional support for the new developments.

References

- [1] Guenther, R., and Myrick, L. *Archiving Web Sites for Preservation and Access: MODS, METS and MINERVA*. Journal of Archival Organization 4, no. 1/2 (2006).
- [2] ISO 14721:2003, CCSDS recommendation: <http://public.ccsds.org/publications/archive/650x0b1.pdf>
- [3] <http://www.kb.nl/dnp/e-depot/dias-en.html>
- [4] <http://kopal.langzeitarchivierung.de/index.php.en>
- [5] Specifications of the Universal Object Model: http://kopal.langzeitarchivierung.de/downloads/kopal_Universal_Object_Format.pdf
- [6] <http://www.loc.gov/standards/mets/>
- [7] Specifications of LMER: <http://nbn-resolving.de/?urn:nbn:de:1111-2005051906>
- [8] <http://netpreserve.org/>
- [9] <http://crawler.archive.org/>
- [10] Draft of the WARC specifications: <http://archive-access.sourceforge.net/warc/>
- [11] <http://archive-access.sourceforge.net/projects/wayback/>
- [12] <http://lcweb2.loc.gov/diglib/lcwa/>
- [13] <http://webcurator.sourceforge.net/>
- [14] <http://webcurator.sourceforge.net/docs/development/WCT%20Project%20SIP%20Specification.doc>

The FRBR-Theoretic Library: The Role of Conceptual Data Modeling in Cultural Heritage Information System Design

Ronald J. Murray

Library of Congress
101 Independence Avenue SE
Washington DC 20012
rmur@loc.gov

Abstract

The use of digital technologies in support of Cultural Heritage missions has highlighted the need to create information modeling systems different from those that are used in conventional business and government. In addition, the practice of data modeling – and especially of the conceptual data modeling that engages cataloging theory and practice – must be urgently brought up to date in order to develop the data models required to represent the desirable characteristics of both print and digital media.

Introduction

The use of digital technologies in support of Cultural Heritage (CH) missions has highlighted the need for information management systems different from those that are used in conventional business, government, and entertainment activities. In particular, a Cultural Heritage institution needs an information system that (a.) supports preservation of and access to both analog and digital content, and (b.) reflects that institution's customary understanding – its view – of the resources it possesses.

The Cultural Heritage community has evolved a number of well-established approaches to the description of resources created in a wide range of media. The hope within that community is that the long-standing theories, practices, and policies that inform the operations of libraries, archives, and museums and provide structure to analog as well as to some digital versions of actual Cultural Heritage content will extend even further to the vast quantities of resources available on the World Wide Web.

The Design of Information Systems Based on Cultural Heritage Concepts

Web-based resources require embodiment, organization, discovery, and access by an electronic information system. The design, implementation, and operation of globally-accessible Cultural Heritage inventory and discovery systems has benefited from collaborative efforts at standardization, with international information technology standards bodies playing a critical role in this effort. However, the crucial data and process modeling steps that lead to the creation of those systems have not received the same level of international attention.

The FRBR Conceptual Model as Exemplar

Since its introduction of the basic concepts underlying the Functional Requirements for Bibliographic Records (FRBR) conceptual model (IFLA 1998), the application of Entity-Relationship Modeling has achieved general acceptance by cataloging theorists. In addition to supporting cataloging theory formation, FRBR was intended to function as a guide for the description of bibliographic materials within and beyond the confines of a library.

However, the literature that details the intervening twenty-one year effort to come to terms theoretically with FRBR (and adapting the model to different media types or to archival records) suggests that deficiencies in or incompatibilities exist with the existing model. These disagreements with the FRBR conceptual model may simply result from problems with *data model quality*: (1.) the current conceptual data model lacks refinement; (2.) the data model reflects an individual modeling style that does not suit the task at hand, and (3.) the FRBR conceptual model's entity, attribute, and relationship definitions reflect mixed or inappropriate data modeling assumptions.

Other explanations for these disagreements are possible. For example, the third data model quality problem above may actually reflect what recent research would identify as an consequence of the complementary stances a data modeler can take relative to the bibliographic "Universe of Discourse" being modeled. More seriously, objections to FRBR may indicate that due to the complexity of the bibliographic universe (and to the numerous ways that interested parties seek to interact with it), there can be no single conceptual data model that will encompass all of the well-established perspectives evolved by archives, libraries, and museums.

The widely discussed and institutionally accepted FRBR conceptual data model can be taken as an indicator of the extent to which the Cultural Heritage community has adopted and the utilized data modeling methodologies that have evolved for purposes of information system design and implementation.

Intent of the Paper

This paper will explore the role that modern data modeling theory and practice has (or has not) played in the development of the FRBR conceptual data model. It also offers examples of modern data modeling approaches

demonstrated with Cultural Heritage subject matter. The analysis is intended to provide guidance to parties attempting to further refine the FRBR conceptual model for theoretical purposes as well as for information system design.

Data Modeling Defined

Data modeling is the step in a database management system design process where things of interest to the enterprise are defined and their relationships delineated. Academic theory and professional educational materials describe data modeling as an interactive process that produces a textual and a diagrammatic representation of an enterprise's information at several levels of abstraction.

Data modeling begins with a review of information system requirements, continues with document reviews and user interviews and model building (in diagram and textual form) with feedback from users. The model may be subject to adjustment to improve performance and is then implemented in a specific implementation technology. Three key data model definitions apply (Hay 2006):

Conceptual Data Model – A description of a portion of an enterprise in terms of the fundamental things of interest to it. They are fundamental in that most things seen by business owners are examples of these.

Logical Data Model – The organization of data for use with a particular data management technology. For relational databases, these are tables and columns; for object-oriented databases, object classes and attributes.

Physical Data Model – The organization of data used to place it in specific storage media. This level refers to “tablespaces” and “cylinders.”

Why a Data Model is Important

Because no database is ever built without a model, the question really becomes whether to model informally or formally, who will be involved, and how much effort will be spent in creating a good design. Data models possess three characteristics that make them essential to system design and implementation:

Leverage – As the data model provides a roadmap for the increasingly technical and implementation-specific representations, programming, etc. that follow, small changes in the data model can have major effects on the system being designed and implemented. A well-designed data model can minimize the need for model changes due to missed requirements and thereby reduce design implementation costs. If the things of interest to the organization are poorly modeled, the database implemented from the model will require more programming effort to input and retrieve data.

Conciseness – Data models provide a compact specification of an information system's requirements and capabilities. Reviewing a data model takes less time than reading a lengthy functional specification document, and makes it easier to obtain an in-depth understanding of the kinds of information that are to be managed.

Data Quality – Problems with data quality (inaccurate data) can often be traced to inconsistency in defining and interpreting data and in implementing enforcement mechanisms for data definitions. Well-defined data model definitions (and enforcement mechanisms) of

dates, addresses, and names preserve the common understanding of what is being recorded and minimizes the need for corrections or workarounds.

What Makes a Good Data Model?

Given that a *designed* data model (as opposed to a faithful description of what is “out there”) is evaluated in terms of how well it meets requirements, a data model quality criterion of must apply. In the absence of a quantitative methodology, Simsion & Witt's criteria are helpful (Simsion & Witt 2005):

Completeness – Does the model support or can it generate the data as specified or implied by the requirements documentation?

Nonredundancy – Does the model preclude the possibility of storing the same fact in more than one place? At the conceptual modeling level, entities that contain the same data would indicate that the model is incomplete and that the model can benefit from the addition of a supertype, where the redundant data can find a home.

Enforcement of Business Rules – How well does the model embody and enforce the rules for handling the data? If data model elements do not allow for specification of all of the conditions that a business imposes on its data, business rules must be elicited and used to further document the model.

Data Reusability – If ways for usefully processing the data are discovered after the model is implemented, is the model flexible enough to permit this without modifying the database? Designing for data independence is very important because data that is organized around a particular application will be harder to adapt when the application changes or is replaced.

Stability and Flexibility – A data model is stable with respect to requirements if a change in requirements does not require changes in the data model. The model is flexible if it can be extended without difficulty to accommodate extensions to existing requirements. Depending on the application environment (e.g. where new media forms are being created, or where cataloging information is being acquired or updated continuously) taking the extra effort to design for stability and flexibility can pay off in reduced data model modification and reduced impact on other implementation levels.

Elegance – Elegance evokes the mathematical sense of the term, where consistency and relative simplicity in describing a model element can be discerned. Elegance in entity definition can be achieved by generalization, for example, when pragmatically compelling entities such as customer, employee, supervisor, security guard, supplier, etc., are generalized into a Party entity that represents these entities as subtypes within a logical, and often hierarchical – structure.

Communication – The ability of the data model to convey its content to technical and non-technical personnel is crucial to determining (a.) whether the model is an accurate representation, and (b.) whether the people who will use or manage the implemented system understand the full implications of the model. Unfamiliar terminology, new concepts, and high levels of complexity tend to render the model less comprehensible to its audience, so

the modeler must organize and present the model with an eye towards maximizing its communicative potential.

Integration – How well does the complete system or system components fit in with what is already there? The ease of difficulty of fit may vary not only with the skill of the designer but also with the novelty of the requirements. For example, library catalog databases that never had to contend with online resources that change on a daily basis may face greater integration challenges than databases where incremental change in resource characteristics is common.

Towards Theory-Guided Design –

A problematic aspect of the data modeling process is that modeling efforts can be undertaken unaware of the *description/design* issue that underlies conceptual data modeling theory and practice. Is the data modeler describing things that are “out there” or is the modeler creating useful data structures that meet specifications? If the Universe of Discourse that is to be represented in the database being modeled contains its *own* unresolved description/design issues (as does the Bibliographic Universe), the result will be a data model where theoretically (or institutionally) compelling model elements become intermixed with elements designed to be useful to programmers and end users. The solution may satisfy the stated requirements rather well, but will please no one.

A good example of an institutionally compelled descriptive or design element (from cataloging theory as well as library tradition) is a hierarchical data structure, which some assert is “the most philosophically interesting of the semantic relationships.” (Svenonius 2001) Notable counterexamples to Svenonius’ interestingness assertion are the network structures that are regularly used to represent a wide range of current theoretical and pragmatic “things of interest” to Communications Theorists, Physicists, Information Scientists, and Political Scientists. (Monge and Contractor 2003; Watts 2003; Csermely 2006) Networks (i.e., graphs, the mathematical structures first described by Euler in 1735) are rarely mentioned in on cataloging theory, nor have they been invoked to describe or define data structures in FRBR.

Consider taking a database *design* perspective to another environment – in the person of a computer programmer at a Physics laboratory where decentralized teams build subatomic particle detectors and conduct research. Network-like structures would be a natural – even unavoidable – part of the intellectual landscape, beginning with a powerful diagrammatic shorthand for describing or hypothesizing particle interactions:¹

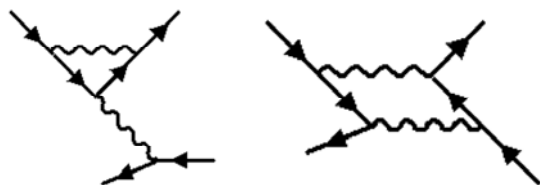


Figure 1: Feynman Diagrams of Electromagnetic Interactions

¹ Feynman Diagrams from <http://www2.slac.stanford.edu/vvc/theory/feynman.html>

The representation of relationships among information resources in that environment might take on an institutionally-compelled network flavor (Berners-Lee and Fischetti 1999), rather than the hierarchical one reinforced by cataloging theory and library institution administrative organization.

A *theory-guided design* solution would be one that evaluates theoretically compelling entities, attributes, and relationships from other fields of endeavor in addition to those originating from cataloging theory. The parties participating in the database management system design process would not be *compelled* to accept these elements, however, just because they have theoretical utility. The design process would benefit from a data model element review that engages a broadened theoretical base to include Social Sciences perspectives – in particular Anthropology, Psychology, Sociology, and Communication Theory.

Modelers initiating a *theory-guided design* strategy can also profit from recently published field research that indicates that professional data modelers depart in significant ways from espoused academic and professional educational teachings, modeling pathways. (Simsion 2007). Data modelers who possessed differing degrees of experience and training were surveyed and also participated in model design experiments.

The respondents were evenly split on the Description vs. Design issue, in spite of the topic never being discussed in the literature. In addition to differences in modeling stance, the data models produced by participants in the data model design experiments demonstrated an effect of experience and personal style on: the number and subtyping of model elements created (fewer used with less experience); the addition of elements and relationships not included in the requirements (more likely with experience and with the *design* stance); and the use of patterns from past modeling projects (more likely with age and the *design* stance).

A Critique of the Current FRBR Model –

FRBR from a modern data modeling perspective – The FRBR conceptual data model as advanced by IFLA raises a number of issues that may be grouped into four categories:

Modeling from legacy systems – Special attention must be paid to the consequences of developing a FRBR conceptual data model that borrows from or must otherwise be made to reflect the structure of legacy logical data models. The danger is that an implementation-specific feature (like a limit set on the number, attribute names or optionality of parties that play *roles* like Author, Editor, Publisher, etc.) will become a requirement to be met in the conceptual data model.

Accommodating legacy systems (in the sense of identifying the functions that were executed by the systems, and understanding the structure of the data in the system) can be made a requirement of an information system. But the design of the new system should not require that identical functions and data structures be created to accomplish this. In the literature, discussions of FRBR model characteristics using patron-oriented legacy displays and scenarios set up by researchers to test hypothe-

ses – have been illustrated using catalog card-like or MARC record-type displays. (Taylor 2007) This indicates that legacy information structures and data presentation strategies continue to dominate designers’ and researchers’ thinking, irrespective of the changes in database design that should be replacing these legacy structures.

Efforts to represent, reason about, and display FRBR bibliographic data should focus more on the data *as it is understood within the new conceptual data model* rather than that of legacy systems. The fact that FRBR conceptual data models contain Many-To-Many network structure means that data modeling efforts should begin with generic network structures for the database design and display, and apply constraints to achieve legacy system hierarchical appearances where unavoidable.

Element use and skill level – The relatively small number of elements and lack of subtyping in the FRBR data model supports Simsion’s finding pertaining to the products of beginning or infrequent modelers. In addition, the model reveals a Cultural Heritage data model documentation bias:

Reliance on the narrative/textual part of the model – By far, the substance of the IFLA FRBR conceptual model specification is textual description (with tables), and only a few diagrams. While these diagrams play a very small role in model documentation and presentation, they are what is used – naturally – to describe the model to the Cultural Heritage community and to the general public. It is difficult to appreciate the overall, emergent, characteristics of the FRBR conceptual data model – especially the more obvious interactions between model elements – from a reading of the text and then attempting to project that wealth of description into the few available diagrams. Especially interesting, but not modeled explicitly, is the means by which the very many neatly subtyped bibliographic relationships defined in the FRBR model text are represented – as attributes, relationships, and possibly even entities.

Simsions’ research revealed that data modeling practitioners – like designers in other fields like engineering, architecture, graphic arts etc., – use diagrams: for their own benefit (contextual placement of model elements with the ability rapidly to modify the model in the face of user feedback, and to detect recurring or out of place patterns); as well to benefit clients (communication of overall model structure and its critical elements).

Missing elements that would improve model communication – Model elements that would make it easier to understand FRBR’s benefits for bibliographic resource discovery are not provided. Also lacking is a distinction between a conceptual data model that is presentable to users (i.e, it is community-specific) vs. one more that is more expressive and accurate for the data modeler and the developers to follow (i.e., it employs data model design conventions and patterns).

Contextualization & Coexistence – The FRBR model at present does not situate its conceptual data model elements within what must be a larger environment of bibliographic and other information resources. The characteristics of information resources of various types, their descriptions, and the roles that institutions can/should play in creating and managing resource descriptions are

therefore not addressed in the model. In a more contextualized model, FRBR and related IFLA data modeling products like name and subject authorities and identifiers – appropriately generalized – have play highly valuable roles to play in the bibliographic universe. These entities benefit from being modeled from a broader perspective.

In the broad context of *Resources*, where *Resource Descriptions* are created to describe the *Resources* that users want to discover and use, it must be stated whether FRBR-based resource descriptions can coexist with other descriptions produced by other institutions or individuals. This issue is not addressed in the present decontextualized model. Description coexistence has significant implications for the placement of resource identifiers, names, and some responsible party roles in the more broadly defined model.

FRBR is not a “Convergent” Conceptual data Model –

A *divergent* conceptual data model is one where entity names, quantities and relationships come from their specific user communities. Similarities in entities and relationships across different data models become difficult to see, and common conventions in data modeling practice are not present (like generalizing entities, standard entity and relationship names, and using patterns). Divergent conceptual data models are very useful however in that they capture a enterprise view that can readily be validated by users.

A *convergent* conceptual data model results when conventions in entity and relationship construction and naming are applied to the divergent model. This step may require the creation of additional elements and relationships based on the modeler’s experience with the structures in the divergent model. (Hay 2005)

A very good indicator that FRBR is a divergent model is the use of community terminology for entity names. Even so, the use of the nondescript prefix “Group” to describe entities should be replaced by meaningful names given these groupings by users. Encouraging users to name data model entities and entity groupings/subtypes is a simple way to induce them to pay more attention to the conceptual data model.

Having commented on FRBR in terms of legacy system issues, element use and institutional preferences for data structures, and on model communication, we can now consider how data modeling can reconcile the desire to build systems that embody well-established intellectually, compelling cataloging concepts with the urge to create data structures that may lack theoretical resonance but get the job done. We propose that the modeling undertaken be guided by – but not be captive to – theory.

Improving on the FRBR Conceptual Data Model

To demonstrate the impact that modern data modeling techniques and conventions can have on increasing the understanding of a CH data modeling effort, the current FRBR conceptual data model – in the form of its diagrammatic representation – will be recast into a different form consistent with modern data modeling practice.

Figure 2 presents a conceptual data models for bibliographic information, names and identifiers, and subjects, respectively. Figure 3 presents a descriptive scenario for

a continuing resource. Space limitations prohibit describing all of the data model elements in the diagram in detail: only a few top-level entities will be described in.

The Larger Context – The entities and relationships defined in FRBR function as descriptions of **Resources**² – analog or digital – that are of interest to one or more persons. For that reason the FRBR entities in this revised conceptual model be defined as **Resource** types. Refer to Figure 1 to clarify the relative placement and connectivity of the entities and relationships to be defined.

Note especially how the compact data model diagram elements expand into a very lengthy set of Business Assertions and comments. Note also that the data model presented in the diagram lies between being a divergent data model and a convergent data model. Some elements are modeled in a conventional fashion, while others remain specific to a Cultural Heritage perspective. This was done deliberately to keep the model somewhat familiar on one hand, but also to highlight design issues on the other.

Design – Beginning with the most general kind of information entity in our Bibliographic Universe, a **Resource** is defined as an information-bearing asset that is drawn upon to accomplish some function.

Commentary – Note the optional (and defined following standard modeling practice) *One-to-Many* relationship at the bottom right of the **Resource** entity. This indicates that a **Resource** may be composed of other **Resources**, of the same or differing **Types**. Defining the ability to “nest” **Resources** at this most basic level makes it possible for the data model elements that are **Resource** subtypes to “inherit” (depending on the conditions we define – we may permit nesting or we may not) the ability to contain sub-**Works**, sub-**Expressions**, sub-**Manifestations**, and sub-**Items**. As we will see, this design decision at the **Resource** level, in combination with judicious Business Rules, resolves a number of issues raised regarding Part/Whole relationships in FRBR entity definition.

Design – We now introduce the four primary **Resource** subtypes:

- **Institutionally Managed Named Resource**
- **Institutionally Managed Named Resource Description**
- **Institutionally Managed Find & Navigate Named Resource Description**
- **Institutionally Managed Find & Navigate Named Resource Assignment**

The characteristics of these different **Resource** types and their relationships with one another will be touched upon briefly. These above **Resource** subtypes make possible sophisticated grouping and referencing FRBR and FRBR-related data model elements.

About the Resources – An **Institutionally Managed Named Resource** is the actual **Resource** that users want to access and use. To efficiently find this **Resource**, an easily accessed *description* of the **Resource** can be consulted. An **Institutionally Managed Named Resource Description** is the means by which users can employ to find/navigate to, identify, select, access and use **Re-**

sources of interest. An **Institutionally Managed Find & Navigate Named Resource Description** is an institutionally managed collection of identifiers, **Resource** names, people, place, concept etc., names, and types of possible relationships between **Resources**. It helps users (thanks to a library catalog some other analog or digital finding aid) to find to the **Resources** they want.

Keeping Track of the Connections – Finally, an **Institutionally Managed Find & Navigate Named Resource Assignment** is a **Resource** that consists of all of the “links” that have been defined between the **Institutionally Managed Find & Navigate Named Resource Descriptions**. Keeping track of the links makes it possible to take shortcuts to **Resources**, and to identify relationships that are not obvious without link information.

Commentary – The FRBR model is not currently defined in **Resource** terms. This makes the relationship between FRBR entity attributes and relationships and its referents in libraries difficult to discern. This decontextualization also makes FRBR entities seem more like *descriptions* derived from theoretical considerations rather than data structures designed to be linked to the actual analog or digital data desired by a user. While the idea of a **Resource** is still rather abstract, the relationship between a **Resource**, a **Resource** description, and the materials on library shelves or server hard drives, and the entries on catalog cards or screen displays is somewhat easier to understand.

An **Institutionally Managed Find & Navigate Named Resource Description**, by its name and by relationships defined in the data model diagram, signals a dependent, helping role with respect to the other two main **Resource** subtypes. The name indicates the role that institutions like libraries, archives, and museums can and do play in making resources (inc. non-bibliographic) easier to access. These institutions standardize names, defining subject headings, identify the persons and organizations etc., that may be sought, and also define the many relationships that exist between all of them.

A **Institutionally Managed Find & Navigate Named Resource Description** is in fact intended to act as a shortcut between what a user knows about – or has on hand in the form of a **Institutionally Managed Named Resource Description** attached to a **Resource** – to the **Resources** elsewhere with the same or similar description(s).

Design – Continuing, a **Resource** may be of one or more **Types**: a **Named Resource** and an **Other Resource**. A **Named Resource** is a **Resource** that is distinguished by the presence of a minimum of three **Institutionally Managed Find & Navigate Named Resource Descriptions**: an **Identifying Authority Resource Description**, a **Responsible Party Resource Description**, and an **Other Relationship Resource Description**.

Defining a Business Rule – It is useful to define the conditions under which a design element can be used. This definition is called a Business Rule, and is considered part of the data model. For a **Resource** to be managed effectively, we will define a Business Rule stating that the **Resource** must possess (a.) one or more unique identifiers, an optional name, and (b.) may optionally be related to another **Resource** in one or more defined ways via an **Other Relationship Resource Description**. This

² In this section, entity names are capitalized and in boldface.

Resource subtype contains the minimum of information required to distinguish it from other **Resources** (via the identifier), and to direct a user from that **Resource** to other **Resources**.

Named Resources are the Basic Units of Discovery and Access – The addition of identifying and relationship information to a basic **Resource** redefines it as a **Named Resource**. A party (identified by its **Responsible Party Resource Description**) can declare itself responsible for a **Named Resource** and then play one or more of several defined roles (e.g., Author, Creator, Publisher, Owner, etc. The institution will decide which *limited set* of possibilities can apply to this subtype) with respect to the **Resource**. A **Resource** that meets this additional responsible party requirement is called a **Managed Named Resource**. The remainder of the data model introduces new types of descriptions that correspond to customary institutional views line those held by librarians, archivists, etc.

What to Do About the FRBR Data Model and Data Modeling in General

Bibliographic information system efforts that rely upon conceptual data modeling can benefit significantly from an infusion of modern conceptual data modeling knowledge, abilities, and skills.

FRBR efforts need to be revisited, with an eye to ensuring that parties currently involved in model development (a.) appreciate the full implications of the original model, and of variations on the model such as the one presented here, and (b.) be prepared to change the model to reflect both improved model understanding and improved techniques for model construction and evaluation.

The talents of professional data modelers should be engaged to monitor data modeling activities taking place in Cultural Heritage institutions. Special effort should be made to have these parties to participate in community-initiated critiques of cataloging theory-based description/design data modeling methods.

A mutually acceptable institution should take leadership in advancing modern data modeling approaches like those introduced here by establishing a Web-accessible data modeling facility accessible to interested Cultural Heritage parties. This facility would (at a minimum) provide or promote training in conceptual data modeling, using well-accepted notations and documentation techniques. In addition the facility should endeavor to:

- Extend modeling activities to other needed areas in the Cultural Heritage realm.
- Design a professional education program and a collegiate curriculum.

This paper has presented a view of current theory and practice of conceptual data modeling, within the context of a unified model of database management system analysis and design. Particular attention was paid to how the FRBR conceptual data model has evolved, and how it differs from this and other alternative models. A conceptual data model that incorporates ongoing IFLA conceptual data modeling initiatives has also been presented and

discussed, along with data model diagrams that address a wide range of content description scenarios.

Analysis of these data models supports a claim that the modeling approach used (*theory guided design*) can employ data modern modeling techniques, while at the same time incorporate the significant intellectual contributions of Cultural Heritage institutions in the realm of resource identification, description, discovery, selection, and access.

Acknowledgements

The author is grateful to Dr. Barbara Tillett for her editorial direction and her monitoring of the data modeling effort presented here.

References

- IFLA 1998. International Federation of Library Associations and Institutions. Functional Requirements for Bibliographic Records. München : K . G. Saur München.
- Simson, Graeme. 2007. Data Modeling: Theory and Practice. Bradley Beach NJ: Technics Press.
- Hay, David. 2006. Data Model Patterns: A Metadata Map. San Francisco: Morgan Kaufman.
- Berners-Lee, Tim and Fischetti, Mark. 1999. Weaving the Web: the original design and ultimate destiny of the World Wide Web by its inventor. San Francisco : Harper San Francisco.
- Simson, Graeme C. and Witt, Graham C. 1994. Data Modeling Essentials. New York: Van Nostrand Reinhold.
- Svenonius, Elaine. 2000. The intellectual foundations of information organization. Cambridge, MA: MIT Press.
- Monge, Peter and Contractor, Noshir. 2003. Theories of Communications Networks. New York: Cambridge University Press.
- Csermely, Peter. 2006. Weak Links: Stabilizers of Complex Systems From Proteins to Social Networks. Heidelberg: Springer.
- Taylor, Arlene. ed. 2007. Understanding FRBR: What It Is and How It Will Affect Our Retrieval Tools. Westport, CT: Libraries Unlimited.
- Hay, David. 2005 Data Model Quality: Where Good Data Begins. The Data Administration Newsletter. <http://www.tdan.com/view-articles/5286>

Towards smart storage for repository preservation services

Steve Hitchcock, David Tarrant, Adrian Brown¹, Ben O'Steen², Neil Jefferies² and Leslie Carr

IAM Group, School of Electronics
and Computer Science, University of
Southampton, SO17 1BJ, UK
{D.Tarrant, S.Hitchcock,
L.Carr}@ecs.soton.ac.uk

¹ The National Archives, Kew,
Richmond, Surrey, TW9 4DU, UK
adrian.brown@nationalarchives.
gov.uk

² Oxford University Library
Services, Systems and Electronic
Resources Service, Osney One
Building, Osney Mead, Oxford OX2
0EW, UK
{Benjamin.Osteen,
neil.jefferies}@sers.ox.ac.uk

Abstract

The move to digital is being accompanied by a huge rise in volumes of (born-digital) content and data. As a result the curation lifecycle has to be redrawn. Processes such as selection and evaluation for preservation have to be driven by automation. Manual processes will not scale, and the traditional signifiers and selection criteria in older formats, such as print publication, are changing. The paper will examine at a conceptual and practical level how preservation intelligence can be built into software-based digital preservation tools and services on the Web and across the network 'cloud' to create 'smart' storage for long-term, continuous data monitoring and management. Some early examples will be presented, focussing on storage management and format risk assessment.

Digital preservation: the big picture

Digital preservation is dealing with a big picture: "A preservation environment manages communication from the past while communicating with the future" (Moore, 2008). In other words, digital preservation might be concerned with any specified digital data for, and at, any specified time. The classic way of dealing with challenges on this scale is to break these down into manageable processes and activities, as digital preservation practitioners have been doing: storage, managing formats, risk assessment, metadata, trust and provenance, all held together and directed by policy.

The advantage digital has over other forms of data is the ability to reconnect, or reintegrate, these components or services, to fulfil the big picture. In this way specified digital content in various locations can be monitored and acted upon by a series of services provided over the Web. Since at the core of any preservation approach is storage, we call this approach 'smart storage' because it combines an underlying passive storage approach with the intelligence provided through the respective services. The key to realising smart storage, as well as building the services, is to enable the services to share information with the digital content sources they may be acting on. This is done through machine-level application programming interfaces (APIs) and protocols, and has become a focus of the work of the JISC-funded Preserv 2 project [Link 1].

Institutional repositories

One of the drivers for the growth of digital content is the Web. The content the project is concerned with is found in digital repositories, specifically in repositories set up by institutions of higher education and research to manage and disseminate their digital intellectual outputs. These institutional repositories (IRs) are a special type of Web site, typically based on some repository software that presents a database of records pointing to the objects deposited. IRs provide varying degrees of moderation on the entry of content, from membership of the institution to some form of light review. Although there are few examples yet of comprehensive policy for these repositories (Hitchcock *et al.* 2007), it is expected the institutions will take a long-term view and that services will be needed to preserve the materials collected by IRs.

The Preserv 2 project is investigating the provision of preservation services for IRs. Rather than viewing itself as a potential service provider, the project is an enabler. It is identifying how machine interfaces can be supported between emerging preservation tools, services, prospective service providers and IRs.

IRs in flux

However, institutional repositories (IRs) are perhaps in a greater state of flux than at any time since their effective inception in 2000 motivated by the emergence of the Open Archives Initiative (OAI). While the number of IRs and the volume of content are growing, there is uncertainty in terms of target content - published papers, theses, research data, teaching materials - policy, rights, even locus of content and responsibility for long-term management.

IRs are developing alongside subject-oriented repositories, some long-established such as the physics Arxiv, while others such as PubMed Central (and its UK counterpart) have been built to fulfil research funder mandates on the deposit and access to research publications. While ostensibly these different types of repository have common aims, to optimise access to the results of research through open access, how they should

align in terms of content deposit policy, sharing and responsibility for long-term management is still an active discussion (American-Scientist-Open-Access-Forum, 2008a).

When planning and costing long-term data management, open access IRs, those targeting deposit of published research papers, in addition need to take account of author agreements with publishers, and of publishers' arrangements for preservation of this content, often in association with national libraries and driven by legal deposit legislation.

Even the infrastructure of IRs is changing. The majority of IRs are built with open source, OAI-compliant software such as DSpace, EPrints and Fedora. The emergence of OAI-ORE (Object Reuse and Exchange, Lagoze and Van de Sompel, 2008) effectively frees the data from being captive in such systems and reemphasises the role of repository software to provide the most effective interfaces for services and activities, such as content deposit, repository management, and dissemination functions such as search, browse and OAI-PMH. The recent emergence of commercial repository services (RSP 2008), from software-specific services to digital library services or more general 'cloud' or network storage services, is likely to further challenge the conventional view of repositories today as a locally-hosted 'box'. It has even been suggested that the 'institutional' role in the IR will resolve to policy, principally to define the target content and mandate its collection for open access, but without specifying the destination of deposits (American-Scientist-Open-Access-Forum, 2008b).

Against this background, where the content and preservation requirements are effectively not yet specified – for IRs we don't know exactly what type of content will be stored, where, and what policy and rights apply to that content and who exercises responsibility for long-term management – it seems appropriate, then, that we consider the big preservation picture and prepare for when the specifics are known and for all eventualities that might prevail at that time.

Towards smart storage

Two characteristics of digital data management, one that applies particularly to digital repositories, are driving approaches towards preservation goals and begin to suggest approaches that we are attempting to identify as smart storage:

- **Scale and economics:** the volume of digital data continues to grow rapidly, while the relative cost of storage decreases, to the extent that services that act on data must be automated rather than require substantive manual intervention, and will demand massive, and probably selectable, storage (Wood 2008)
- **Interoperability:** the viability of IRs is predicated on interoperability provided by the

OAI Protocol for Metadata Harvesting (OAI-PMH), to enable the aggregated contents of repositories to be searched and viewed globally rather than just locally. We now seek to exploit interoperability in the wider context of what is more clearly recognised as the operative Web architecture, known as Representational State Transfer, or RESTful, and is the basis of many Web 2.0 applications that expose and share data

Open storage

In terms of content and data, IRs are characterised by openness: the most widely used repository softwares are open source, and the content in IRs is largely open access. From the outset IRs have been 'open archives' having adopted the OAI-PMH to share data with e.g. discovery services. Now OAI has been extended to support object reuse and exchange, which enables the easy movement of data between different types of repository software, giving substance to the concept of 'open repositories'. More recently we have seen the emergence of large-scale storage devices based on open source software, leading to the term 'open storage'.

Using open storage averts the need for a repository layer to access first-class objects – these are objects that can be addressed directly – where first-class objects include metadata files which point to other first-class objects (such as an ORE representation). We can now begin to realize situations where an institution can exploit the resulting flexibility of repository services and storage: multiple repository softwares can run over a single set of digital objects; in turn these digital objects can be distributed and/or replicated over many open storage platforms.

Being able to select storage enables platforms with error checking and correction functions to be chosen, such as parity (as found in RAID disc array systems), bit checking – a method to verify that data bits have not become corrupted or “switched” – self-recovery and easy expansion. Ordinarily, for economic reasons repositories might not have use of these more resilient storage platforms, but they may become viable for preservation services aimed at multiple repositories.

Early adopters of open storage include Sun Microsystems, which is developing large-scale open source storage platforms, including the STK5800 (codenamed Honeycomb). By focusing on object storage rather than file storage the Honeycomb server provides a resilient storage mechanism with a built-in metadata layer. The metadata layer provides a key component in open storage where objects are given an identifier. For repositories using open storage, there are two scenarios:

1. The repository creates a unique identifier (UID) and URL for an object and the storage platform has to know how to retrieve this object given this identifier.

2. The storage platform creates the UID and/or URL and passes this to the repository on successful creation of the object.

We envisage that both will need to be supported; the first is suited for offline storage mechanisms, whereas the second can be used for cloud and Web 2.0 storage mechanisms.

Aligning with the Web architecture

Three architectural bases of the Web are identification, interaction and formats (Jacobs and Walsh, 2004). It is notable how Web 2.0 applications are designed to be more consistent with the Web architecture than previous-generation Web applications. ORE, for example, with its use of URIs for aggregate resource maps as well as individual objects, opens up new forms of interaction for repository data and extends OAI to conform with Web architectural principles.

We can recognize the growing prevalence of these features, particularly in the number of available APIs. Major services on the Web, such as Google Maps, deploy their own simple APIs. An example within the repository community is SWORD (Simple Web-service Offering Repository Deposit), and open storage platforms such as Sun's STK5800 and the Amazon Simple Storage Service (S3) can similarly be accessed by simple, if different, APIs. To take advantage of open storage, repositories have to be able to talk to these services through these APIs.

An extra feature of STK5800 is Storage Beans, programming code that enables developers to create applications to run on the platform. This is helpful when objects and data need to be manipulated without removing them from the archive.

There is a temptation to try and create standards for methods of communication between applications, especially as in the cases below where the range of potential applications that we may want to work with can be identified. At this stage it appears inevitable that we will have to be adaptable and work with the continuing proliferation of APIs.

Application examples

Storage management

Open repository platforms, which are essentially a set of user and machine interfaces to a built-in storage or database application, are starting to abstract their storage layers to provide flexibility in choice of storage approaches. Increasingly repositories are seen, from a technical angle, as part of a data flow, rather than simply a data destination, and the input and output of data from repositories is supported by applications or interfaces called 'plugins', which can be developed and shared independently without having to modify the core repository software. Typical examples include import

and export of different metadata and reference formats, transfer of XML records, RSS feeds, or data for timelines (Figure 1). EPrints, from version 3.0, is a prominent example of this approach.

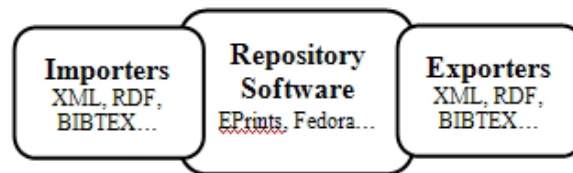


Figure 1: Plugin applications for EPrints prepare data formats for import to, export from, repositories

Adopting the same approach, Preserv 2 is working with the JISC Common Repository/Resource Interface Group (CRIG) and the EPrints technical team to develop a set of expandable plugins to interface EPrints with many types of storage including online and open storage platforms. In addition, EPrints provides a scriptable Storage Controller allowing more than one plug-in to be used to send objects to different storage destinations (Figure 2) based, for example, on the properties of the object or on related metadata. By allowing more than one plugin to be used concurrently it is possible for a plugin to be used specifically for the purposes of long-term preservation services.

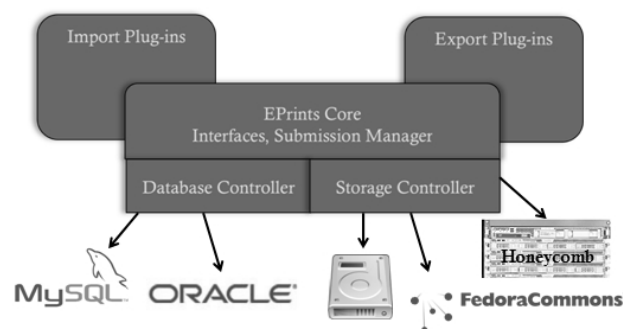


Figure 2: Storage controller, as implemented for EPrints software, enables selected plugins to interface with chosen storage

EPrints is not the only platform developing this sort of architecture. The Akubra project is looking at pluggable low-level storage for Fedora repository software.

Format services

If storage is intended to be a 'passive' preservation approach, in that the aim is to keep the object unchanged, a more active approach is required to ensure that an object remains usable. This requires identification of the format of a digital object and an assessment of the risk posed by that format.

Digital objects are produced, in one form or another, using application programs such as word processors and other tools. These objects are encoded with information to represent characters, layout and other features. The

rules of the encoding are defined by the chosen format of the object. Applications are often closely tied to formats. If applications and formats can change over time, it follows that some risk becoming obsolete – if an application is superseded or becomes unavailable it may not be possible to open objects that were created with that application. This is why formats are a primary focus for preservation actions. The risk to a format can be monitored and might depend on several factors, such as the status of the originating application, or the availability of other tools or viewers capable of opening the format. In some cases objects in formats found to be at-risk may be transformed, or migrated, to alternative formats.

It can be seen from this description that preservation methods affecting formats can be classified in three stages:

- Format identification and characterization (which format?)
- Preservation planning and technology watch (format risk and implications)
- Preservation action, migration, etc. (what to do with the format)

Format-based services tend to be *ad hoc* processes for which some tools are available but which few systems use in a coordinated manner. Currently none of the repository platforms offer support for these tasks beyond basic file format identification using the file extension. Such preservation services can either be performed at the repository management level, or by a trusted third-party service provider. Preserv 2 is working on supporting format services in the cloud alongside open storage, transforming open storage into smart storage. The types of preservation services we are addressing here include file format identification (more than simple extension), risk analysis, and location and invocation of migration tools. All of these require interaction with the repository and access to repository policies. This introduces the need for messaging between the service and the repository, which we address in relation to the services outlined.

Our starting point for this work on smart storage architectures takes existing preservation tools such as PRONOM-DROID (PRONOM [2] is an online registry of technical information, such as file format signatures; DROID [3] is a downloadable file format identification tool that applies these signatures) from The National Archives (UK). In the first phase of Preserv, DROID was implemented as part of a Web service, automatically uploading files from repositories for classification (Brody *et al.* 2007). This uses a lot of bandwidth for large objects, however, and DROID can also become quite processor-intensive. Thus placing this tool alongside storage can decrease the load and bandwidth requirement on the repository while providing most benefit.

Figure 3 shows the implementation of DROID within a smart storage environment. DROID is unchanged from the version distributed by TNA, but three interfaces enable it to interact with an open storage platform and a repository, in this case based on EPrints, which has minor schema changes so that it can accept the metadata generated by DROID.

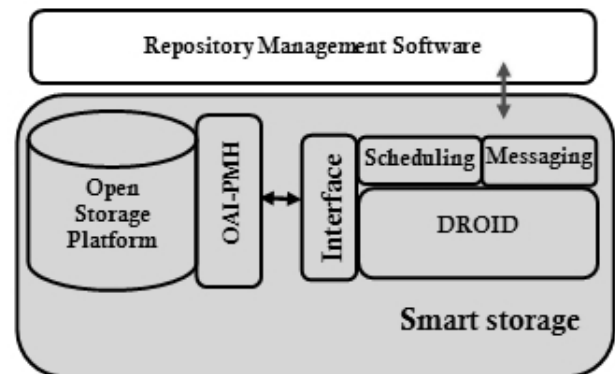


Figure 3: DROID (Digital Record Object Identification) within a smart storage arrangement

The first interface invoked is scheduling, which controls when an update needs to be performed. Preserv 2 has developed a scheduling service based on the Apple iCal calendar format. This interface can thus be controlled directly by the repository by a default repeating event or by a synchronized desktop calendar client. This provides a powerful scheduling service with many clients already available that can read and interpret the files so that both past and future events can be reviewed. In this case the controller around DROID will write the output log into the scheduled event in a log file-type format.

It is anticipated the scheduler will invoke actions based on the results of scanning by DROID allied to decision-making tools that use intelligence from planning and technology watch tools, such as the Plato [4] preservation planning tool from the EC-funded Planets [5] project.

An OAI-PMH interface to open storage discovers the latest objects to have been deposited and which are ready for format classification. Using OAI-PMH is one example of an interface to DROID that can perform this function, but it could also be performed by simpler RSS or Atom-based methods. This interface has since been expanded, again alongside work being done with EPrints, to allow export of OAI-ORE resource maps in both RDF and Atom formats (using the new ORE `rem_rdf` and `rem_atom` datatypes, respectively).

Once new content is discovered a simple controller (not shown in Figure 3) feeds relevant information to DROID, which performs the classifications. At this stage the scheduler is updated and the results are fed to any subscribers, currently by pushing into EPrints.

As a final note on Figure 3 it can be seen that these services and interfaces have been encapsulated within a

smart storage box. Each service has been implemented as Java code and each is able to run alongside the services that are managing the storage API and bit checking.

This implementation provides an early indication of how a decoupled service will need to interface with a range of services and repository management softwares. The simplest method encourages the use of XML and/or RDF for call and callback to and from services. If callback is to happen dynamically between the repository and smart storage, a level of trust needs to be established with this service, and simple HTTP authentication will be required in future releases. A key feature is that all services use RESTful methods for communicating, thus maintaining consistency with the Web architecture, enabling easy plug-ability of new or existing services to a repository.

Further work

Further services are being developed that will be able to interface with representation information registries (Brown 2008) such as PRONOM, which expose information for use by digital preservation services. PRONOM is being expanded as part of Preserv 2 and the EC-funded Planets project to include authoritative information on format risk. Alongside format information a user/agent will then be able to request a risk score relating to a format. This score will be calculated based on several factors each of which has a number of step-based scoring levels, e.g. number of tools available to edit the format.

The Plato preservation tool from the Planets project offers another, in this case user-directed, way of classifying format risks based on specified requirements. The importance of such an approach is that it can take into account the significant properties or particular use cases of a digital object (Knight 2008). Properties of an object that might be considered significant can vary depending who specifies them. Creators, repository managers, research funders in the case of scholarly work, and preservation service providers, can each bring a different view to the features of a digital object that have to be maintained to serve the original purpose.

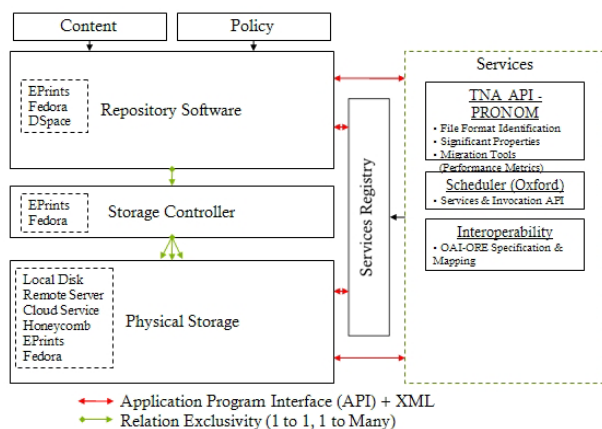


Figure 4: Storage-services based model of Preserv 2 development programme

A more complete picture of how the smart storage approach outlined here fits into the broader programme of Preserv 2 is shown in Figure 4.

Summary

We can place our concept of smart storage within a range of storage approaches and identify a progression:

1. binary stream
2. file system - need to store multiple streams with permissions
3. content addressable - adds content validation and object identifiers, metadata required to locate an object
4. open - adds error correction and recovery, places processing close to storage, solves some bandwidth problems
5. smart - opens up the close-to-storage approach for application development, transition to 'cloud' storage

We also begin to see how smart storage can address the storage problems we encounter:

1. "Billion file" issue - technical scalability of file systems (Wood 2008)
2. Retrieval/indexing - how to locate an item
 - a simple hierarchy is no longer sufficient (RDF maps needed)
 - expectation of Google-style accessibility
 - indexes can themselves require significant storage/processing
3. File integrity - checking, validation, recovery
 - backup as an approach does not scale
 - soft errors become significant
 - bandwidth limits speed of checking, recovery and replication
4. Security/preservation - need for more extensive metadata
 - layered, orthogonal functions over basic storage
5. Application scalability/longevity
 - need to decouple components (Web services or plugins approach, for example)
 - but some functions are bandwidth-hungry, so we need balanced storage/processing at the bottom level
 - use of platform independence (Java, standard APIs) so a "storage bean" can migrate across nodes
 - tightly-coupled Honeycomb is not the only approach, SRB/IRSDS is looser
 - with OAI-ORE objects can migrate too
 - very "cloud"-y

- heterogeneous environment - storage policy for different applications/media types, delivery modes

The emergence of this preliminary but flexible framework for managing data from repositories, and the convergence of preservation tools and services, provides the opportunity to reexamine the curation lifecycle, which is being challenged by sharply growing volumes of digital data. The trick will be to identify those traditional approaches that continue to have value, and to adapt and reposition these within the new framework, typically within software. Openness, in its various forms, the ability to move data freely and easily, needs to be supplemented by decision-making that can be automated based on the supplied intelligence and information. In this way, open storage can become 'smarter'.

References

- American-Scientist-Open-Access-Forum, 2008a, Convergent IR Deposit Mandates vs. Divergent CR Deposit Mandates, from 25 July 2008
<http://tiny.cc/yM017>
 or see <http://amsci-forum.amsci.org/archives/American-Scientist-Open-Access-Forum.html>
- American-Scientist-Open-Access-Forum, 2008b, The OA Deposit-Fee Kerfuffle: APA's Not Responsible; NIH Is, see Harnad, S., July 17, and Hitchcock, S., July 18
<http://tiny.cc/Y1wDl>
 or see <http://amsci-forum.amsci.org/archives/American-Scientist-Open-Access-Forum.html>
- Brody, T., Carr, L., Hey, J. M. N., Brown, A. and Hitchcock, S., 2007, PRONOM-ROAR: Adding Format Profiles to a Repository Registry to Inform Preservation Services, *International Journal of Digital Curation*, Vol. 2, No. 2, November
<http://www.ijdc.net/ijdc/article/view/53/47>
- Brown, A., 2008, Representation Information Registries, Planets project, White Paper, 29 January
http://www.planets-project.eu/docs/reports/Planets_PC3-D7_RepInformationRegistries.pdf
- Hitchcock, S., Brody, T., Hey J. M. N. and Carr, L., 2007, Survey of repository preservation policy and activity, Preserv project, January
<http://preserv.eprints.org/papers/survey/survey-results.html>
- Jacobs, I. and Walsh, N. (eds), 2004, Architecture of the World Wide Web, Volume One, W3C Recommendation, 15 December
<http://www.w3.org/TR/webarch/>
- Knight, G., 2008, Framework for the definition of significant properties, version: V1, AHDS, InSPECT Project Document, 5 February
<http://www.significantproperties.org.uk/documents/wp33-propertiesreport-v1.pdf>

Lagoze, C. and Van de Sompel, H. (eds), 2008, ORE Specification and User Guide - Table of Contents, 2 June 2008

<http://www.openarchives.org/ore/toc>

Moore, R., 2008, Towards a Theory of Digital Preservation, *International Journal of Digital Curation*, Vol. 3, No. 1

<http://www.ijdc.net/ijdc/article/view/63/82>

RSP, 2008, Commercial Repository Solutions, Repositories Support Project

<http://www.rsp.ac.uk/pubs/briefingpapers-docs/technical-commercial-solutions.pdf>

Wood, C., 2008, The Billion File Problem And other archive issues, *Sun Preservation and Archiving Special Interest Group (PASIG) meeting*, San Francisco, May 27-29, 2008

http://events-at-sun.com/pasig_spring/presentations/ChrisWood_MassiveArchive.pdf

Links

- [1] Preserv 2 project <http://preserv.eprints.org/>
- [2] Online registry of technical information, PRONOM <http://www.nationalarchives.gov.uk/pronom/>
- [3] DROID (Digital Record Object Identification) <http://droid.sourceforge.net/wiki/index.php/Introduction>
- [4] Plato - Preservation Planning Tool <http://www.ifs.tuwien.ac.at/dp/plato/>
- [5] Planets - Preservation and Long-term Access through NETworked Services <http://www.planets-project.eu/>

Repository and Preservation Storage Architecture

Keith Rajecki

Sun Microsystems, Inc.
15 Network Circle
Menlo Park, CA 94025 USA

keith.rajecki@sun.com

Abstract

While the Open Archive Information System (OAIS) model has become the de facto standard for preservation archives, the design and implementation of a repository or reliable long term archive lacks adopted technology standards and design best practices. This paper is intended to provide guidelines and recommendations for standards implementation and best practices for a viable, cost effective, and reliable repository and preservation storage architecture. This architecture is based on a combination of open source and commercially supported software and systems.

Although several operating systems currently exist, the logical choice for an archive storage system is an open source operating system, of which there are two primary choices today: Linux and Solaris. There are many varieties of Linux available and supported by nearly all system manufacturers. The Solaris Operating System is freely downloadable from Sun Microsystems. Many variants of the Linux operating system and Solaris are available with support on a fee base.

The Hierarchical Storage System, or HSM, is a key software element of the archive. The HSM provides one of the key components that contributes to reliability by through data integrity checks and automated file migration. The HSM provides the ability to automate making multiples copies of files, auditing files for errors based on checksum, rejecting bad copies of files and making new copies based on the results of those audits. The HSM also provides the ability to read in an older file format and write-out a new file format thus migrating the format and application information required to ensure archival integrity of the stored content. The automation of these functions provides for improved performance and reduced operating costs.

The Sun StorageTek Storage Archive Manager (SAM) software provides the core functionality of the recommended preservation storage architecture. SAM provides policy based data classification and placement across a multitude of storage devices from high speed disk, low cost disk, or tape. SAM also simplifies data management by providing centralized meta-data. SAM is a self-protecting file system with continuous file integrity checks.

The digital content archive provides the content repository (or digital vault) within Sun's award-winning Digital Asset Management Reference Architecture (DAM RA). DAM RA

enables digital workflow and the content archive provides permanent access to digital content files.

With SAM software, the files are stored, tracked, and retrieved based on the archival requirements. Files are seamlessly and transparently available to other services. SAM software creates virtually limitless capacity. Its scalability allows for continual growth throughout the archive with support for all data types. The policy based SAM software stores and manages data for compliance and non-compliance archives using a tiered storage approach with integrated disk and tape into a seamless storage solution, SAM software simplifies the archive storage. Allows you to automate data management policies based on file attributes. You can manage data according to the storage and access requirements of each user on the system and decide how data is grouped, copied, and accessed based on the needs of the application and the users. Helps you maximize return on investments by storing data on the media type appropriate for the life cycle of the data and simplifying system administration.

Sun Open Storage solutions provide the systems built with an open architecture using industry-standard hardware and open-source software. This open architecture allows the most flexible selection of the hardware and software components to best meet storage requirements. In a closed storage environment, all the components of a closed system must come from the vendor. Customers are locked into buying disk drives, controllers, and proprietary software features from a single vendor at premium prices and typically cannot add their own drives or software to improve functionality or reduce the cost of the closed system. Long term preservation is directly dependant on the long term viability of the software components. Open source solutions offer the most viable long term option with open access and community based development and support.

Repositories and Preservation Storage Architecture

The Repository and Preservation Storage Architecture illustrates the integration of Sun software into the implementation of digital repositories and preservation archiving software on Sun systems. This architecture delivers extreme levels of availability and offers proven enterprise-class scalability. The architecture includes specific recommendations for hardware and software components that can help improve manageability, operational performance and efficient use of storage

infrastructure.

Guidelines and Recommendations on Building a Digital Repository and Preservation Archive

The first step to building a repository and preservation storage architecture is the assessment of the business processes and defining the goals of your repository and preservation archive. Incorporating the business processes into your architectural design is crucial to the overall success of the long term archive. Documenting your organizations policies and procedures including data types, length of archive, access methods, maintenance activities, and technical specifications will increase the probability your archive architecture will meet the business requirements.

A reliable long term archive is also dependant on the software components being open and supporting interoperability. Storing, searching, and retrieving data is not sufficient criteria for a successful long term archive. A long term archive should incorporate open source standards based software to ensure future support.

The overall storage system architecture addresses the physical storage components and processes for long-term preservation. Key components to address when architecting your long-term archive are security, storage, and application interoperability. The security layer focuses on the data access in order to ensure integrity and privacy. Storage addresses the placement of the objects within the various hardware components based on retention policies. Application interoperability is the systems and applications ability to be backward compatible as well as the ability to support expanded system functionality.

When designing your repository or preservation archive system it is important to understand the needs of the users of the system. Users are not limited to those who will be accessing the repository or archive looking for objects, but includes those who will be ingesting objects as well. Your users may consist of students, faculty, researcher, or event the general public. Each of which may have different access needs. These needs will influence the server requirements of your access tier as well as the performance requirements of your search and data retrieval. You must be able to define your acceptable levels of retrieval response times in order to ensure your objects are being stored on the most appropriate storage device. High speed disk systems will provide you with faster data access compared to tape library that may need to search and mount media prior to retrieval.

Funding is also an important consideration when planning your repository or preservation archive system. You must consider the operating and upgrade cycles of your architecture in addition to the initial acquisition costs. This will prevent you from implementing a solution that is either too costly to maintain or requires drastic re-architecture as a result of the growth of the repository. This architecture takes advantage of low

cost storage combined with open standards that lower your total cost of ownership.

This architecture supports a wide variety of content types. When planning your repository or preservation archive, you should consider the various content types you will be required to support. You may want to begin evaluating and planning different preservation policies for different content types. Not all content has the same preservation requirements or value. Flexibility of the tiered storage architecture allows you to expand and contract your individual storage tiers independently as your content storage requirements evolve. Here are a few examples of some of the content type you may be consider digitizing, ingesting, and preserving in your repository:

- Manuscripts
- Books
- Newspapers
- Music, Interviews, and Video
- Web Documents and Content
- Scientific and Research Data
- Government Documents
- Images
- eJournals
- Maps

In addition to understanding your digital object types, you also want to consider the size of those objects as well as the total size of the repository. This will also allow you to forecast the growth rate of your digital repository in terms of the number of objects, object size, replication of objects, and total storage capacity. You will also want to establish and adhere to standard file formats when storing your digital objects such as tiff, jpg, or txt. It will be important that these file formats can be read by the applications that are available in the future when they are accessed from the repository or archive..

Repository Solutions

The term repository is widely debated by some. For the purposes of this solution architecture, repository refers to the system by which objects are stored for preservation archiving. There are a number of viable repository solutions available that provide the capability to store, manage, re-use and curate digital materials. Repository solutions support a multitude of functions and can be internally developed or extended. These repository solutions were highlighted for their ability to integrate into a tiered storage architecture and their support for interoperability. The repositories must be sustainable and supportable in order for the underlying storage system to operate.

Fedora

Fedora is developed by the Fedora Commons non-profit organization as a platform for providing sustainable technologies to create, manage, publish, share and preserve digital content as a basis for intellectual,

organizational, scientific and cultural heritage. Fedora is open source software built around a robust integrated repository-centered platform that enables the storage, access and management of virtually any kind of digital content. Content in Fedora can easily be accessed from the Web or by almost any software applications using available extensible application programming interfaces (API's). The connections between content items can be captured and stored in Fedora as semantic relationships describing both the linkage and its meaning.

Fedora is the first open source repository designed to work as part of an extensible framework of service components. This allows you to seamlessly incorporate Fedora into your organization's existing infrastructure. This extensible framework also allows Fedora to support trusted, secure organizational repository needs while supporting rapidly changing Web services applications. Fedora's standards-based framework can incorporate the latest technology while keeping the content safe and accessible. Using this framework, you can easily add innovative technologies as services or plug-ins without compromising the trusted core.

DSpace

DSpace is an open source digital repository system that allows researchers to capture, store, index, preserve and redistribute digital data in virtually any format. More than 300 institutions worldwide use DSpace as their digital repository. DSpace provides organizations with an easy to use end-to-end solution for managing and providing permanent access to their digital works. DSpace was originally developed as a joint effort between MIT Libraries and Hewlett-Packard (HP). It is freely available to all commercial and non-commercial organizations under the BSD open source license. DSpace is designed to work out of the box and yet it also provides the flexibility to be easily customized to meet an institution's unique needs. DSpace Manakin provides a modular user interface layer, enabling institutions to design a unique look-and-feel that can be different for each community, collection and item across the repository. Manakin also allows the user interface to extend outside of DSpace into an existing Web presence. DSpace supports multiple types of storage devices through a lightweight storage API. The storage layer currently provides support for local file systems, Storage Resource Broker (SRB), Amazon S3, or Sun SAM/QFS. New storage devices or approaches can be quickly integrated using the existing storage API's.

EPrints

EPrints is an open source software package for building open access repositories that are compliant with the Open Archives Initiative Protocol for Metadata Harvesting. It shares many of the features commonly seen in Document Management systems, but is primarily used for institutional repositories and scientific journals. EPrints was developed at the University of Southampton School of Electronics and Computer Science and is released under a GPL license. EPrints is a Web and command-line application based on

the LAMP architecture but has been ported and optimized for Solaris. Version 3 of the software introduced a (Perl-based) plugin architecture for importing and exporting data, as well as converting objects (for search engine indexing) and user interface widgets.

VTLS Inc. Vital

VITAL is a commercial institutional repository solution from VTLS Inc. designed for universities, libraries, museums, archives and information centers. Built on Fedora™, this software is designed to simplify the development of digital object repositories and to provide seamless online search and retrieval of information for administrative staff, contributing faculty and end-users. VITAL provides all types of institutions a way to broaden access to valuable resources that were once only available at a single location and to a finite number of patrons. By eliminating the traditional limitations information seekers encounter, this technology grants access to materials for all authorized end-users, from professional researchers to recreational learners. Vital is a perfect solution for organizations looking for a commercially supported alternative to open source applications.

Storage Architecture Components

Whether you are building a repository for managing institutional content, to preserve historical material, to store data for business compliance, or meet evolving business needs, a tiered storage architecture can provide you with the most reliable and cost effective solution. If architected incorrectly, ingest, searching, and preservation can be time consuming and costly. Traditional tape only archival methods simply can not meet the access requirements of many of today's repositories and long term archives. Likewise, storing all the data on disk requires greater administration and is more costly. The proposed architecture provides a proven solution with a balance between disk and tape storage hardware to support long term archiving.

Storage Archive Manager (SAM/QFS)

The Sun StorageTek Storage Archive Manager (SAM) software provides the core functionality of the recommended archive storage architecture. SAM provides policy based data classification and placement across a multitude of tiered storage devices from high speed disk, low cost disk, or tape. SAM also simplifies data management by providing centralized metadata. SAM is a self-protecting file system with continuous file integrity checks.

Sun Storage Archive Manager addresses compliance by applying policies to files, copying and moving files based on those policies and maintaining audit information on files. SAM indexes files for searchability and writes multiple copies to specific media based on the compliance retention policies.

Designed to help address the most stringent requirements

for electronic storage media retention and protection, Sun StorageTek Compliance Archiving Software provides compliance-enabling features for authenticity, integrity, ready access, and security.

Key Benefits of StorageTek Compliance Archiving Software

- Enforces retention policies at the storage level
- Software-controlled disks implement non-rewritable and non-erasable files
- Offers a cost-effective alternative to optical or tape archives
- Provides flexible Fibre Channel or SATA configurations

StorageTek Compliance Archiving software implements write-once read-many (WORM) files that are nonrewritable and nonerasable. Robust security features such as audit logs, user authentication, and access controls, combine to help safeguard the integrity of the digital information. In addition, the critical metadata attributes cannot be changed.

The Hierarchical Storage System, or HSM, is a key software element of the archive. The HSM provides one of the key components that contributes to reliability through data integrity checks and automated file migration. The HSM provides the ability to automate making multiples copies of files, auditing files for errors based on checksum, rejecting bad copies of files and making new copies based on the results of those audits. The HSM also provides the ability to read in an older file format and write-out a new file format thus migrating the format and application information required to ensure archival integrity of the stored content. The automation of these functions provides for improved performance and reduced operating costs.

Sun Fire X4500 Server

The Sun Fire X4500 Server provides a single platform for both applications and data, with enterprise server reliability features and extremely high data throughput rates. The integration of server and storage technologies, makes this an ideal platform for an inexpensive clustered storage tier. The Sun Fire X4500 Server delivers the remarkable performance of a four-way x64 server and the highest storage density available, with up to 48 TB in 4U of rack space. This system also delivers incredibly high data throughput for about half the cost of traditional solutions.

Sun Customer Ready Infinite Archive System

The Sun Customer Ready Infinite Archive System provides a pre-installed and configured storage solution for digital repository and preservation archiving. The Infinite Archive solution scales easily providing petabyte scalability. The Sun Customer Ready Infinite Archive System provides a three tier storage system consisting of the following components.

- Working Data Set, Online, on fast Fibre Channel (FC) Storage (Sun StorageTek 6140 Array)

- First Level Archive, Midline, high capacity SATA storage (Sun StorageTek 6140 Array)
- Second Level Archive, Nearline, high-performance tape storage (Sun StorageTek SL500 Modular Library System)
- Remote Archive provides a further level of archiving, with remote off-site storage of archived tapes

The Infinite Archive System takes advantage of Sun SAM/QFS software to manage the placement and retention of the data to ensure the most cost effective use of your storage resources.

Sun StorageTek 6140 array

The Sun StorageTek 6140 array is the perfect blend of performance, high availability, and reliability. The StorageTek 6140 array architecture scales to 112TB per system including the non-disruptive addition of capacity and volumes, RAID and segment size migration, and switched technology with point-to-point connections. All components in the array's data path are redundant and eliminate any single point of failure. If one component fails, the StorageTek 6140 array automatically fails-over to the alternate component, ensuring continuous uptime and uninterrupted data availability. Every component in the StorageTek 6140 array (from the disk drives to the midplane) is hot-swappable. Hot spares in every storage tray of the StorageTek 6140 array ensures high availability. Hot-spare drives can be allocated from unused drives and are always powered up and available as a spare to any virtual disk in any tray. Each array controller has two power supplies, each with its own battery backup system providing redundant power.

The StorageTek 6140 array easily adapts to change without disrupting existing applications. Compatible storage modules enable non-disruptive system upgrades and data-in-place migration of assets. The compatible and common array management across the entire Sun StorageTek Series 6000 product line protects your investment in management tools, training, and forklift upgrades.

Sun CoolThreads Servers

Sun systems with CoolThreads technology deliver breakthrough performance with dramatic space and power efficiency. Sun CoolThreads Servers are powered by the UltraSPARC T2 or T2 Plus processor, the industry's first "system on a chip" packing the most cores and threads of any general-purpose processor available. These unique servers offer energy efficiency and high performance for vertical and horizontal scalability. The Sun SPARC Enterprise T5140 and T5240 servers utilize the UltraSPARC® T2 Plus processor, which adds multisoocket capabilities to the successful UltraSPARC T2 processor. These servers are ideal for meeting the demands of ingest, web services, and metadata management.

Sun StorageTek Modular Library System

The Sun StorageTek Modular Library Systems are the

most scalable solutions on the market with up to 56 petabytes and 70,000 tape slots. This makes them the ideal platform for tape archives for off-line or dark archives. The Sun StorageTek Modular Library is complemented by the Sun StorageTek VTL Plus or Sun StorageTek VTL Value virtual solutions, which integrate seamlessly with physical tape. As a result, you gain a no-compromise solution that balances the performance, reliability, and ease of management of VTL to enable tape consolidation with the low cost, cartridge removability, and long-term retention capabilities. This tiered storage solution is managed by policies on the VTL, so the overall solution reduces your labor costs for virtual and physical tape management.

The StorageTek Modular Library Systems provide greater levels of reliability ensuring access to your data. The robotic mechanism maintains reliability regardless of the number of expansion modules and helps to increase the stability and predictability of backups. Redundant, hot-swappable components, such as power supplies and fans, minimize disruption. An advanced digital vision system automatically calibrates the library to reduce wear and tear on the cartridge, drive, and robot. Dynamic worldwide naming and firmware code uploads eliminate single points of failure.

Sun Identity Management Suite

The Sun Identity Management Suite is a key component to ensuring the security and data integrity of the digital repository and preservation archiving solution. Identity Manager provides a comprehensive user provisioning and identity auditing for efficiently and securely managing identity profiles and permissions while Sun Directory Server Enterprise Edition provides a secure, highly available, scalable, and easy-to-manage directory infrastructure that effectively manages identities in this growing and dynamic environment.

Solaris

Although several operating systems currently exist, the logical choice for an archive storage system is an open source operating system, of which there are two primary choices today: Linux and Solaris. There are many varieties of Linux available and supported by nearly all system manufacturers.

The Solaris Operating System is freely downloadable from Sun Microsystems and provides a number of technical advantages from file system support to security and supportability. The Solaris ZFS offers a dramatic advance in data management with an innovative approach to data integrity, performance improvements, and integration of file system and volume management capabilities. Solaris Dynamic Tracing (DTrace) allows you to analyze, debug, and optimize your systems and applications.

The Solaris OS also offers binary compatibility within each Sun server line, whether based on UltraSPARC®, AMD Opteron, or Intel Xeon processors. As a result, all Sun servers running the Solaris 10 OS provide powerful

features that can help reduce cost, complexity, and risk. Many variants of the Linux operating system and Solaris are available with support on a fee base.

Conclusion

A tiered storage architecture provides the most cost effective solution for object repositories and long-term archives while supporting scalability. The extent at which those storage tiers are deployed is dependant on the access patterns and archival policies. Although this architecture is not intended to cover all business requirements, it can be applied in a modular approach to address specific business requirements where one or more tiers may not be feasible due to business or technical requirements.

Repository and Preservation Storage Architecture Key Benefits

The architecture identifies key system components and processes that are required to achieve high service levels and scalability. It provides the following major benefits to educational institutions:

- Higher service levels — The architecture is designed to optimize service levels with redundant components and automated failover using storage virtualization and cluster technologies.
- Reduced cost — Virtualization technologies enable consolidated solutions with higher resource utilization and tiered storage helps customers avoid overprovisioning or underprovisioning their systems. Best practices for management can also reduce the cost of maintaining the solution environment.
- Faster time to delivery — Accelerates deployment by providing proven and tested configurations with simplified installation to be up and running almost immediately.
- Reduced risk — Validated hardware and software configurations greatly reduce the risk of unforeseen problems in a production implementation.

References

Consultative Committee for Space Data Systems (2002). "Reference Model for an Open Archival Information System (OAIS)". CCSDS 650.0-R-1 – Blue Book. Available at: <http://ssdoo.gsfc.nasa.gov/nost/wwwclassic/documents/pdf/CCSDS-650.0-B-1.pdf>

Repository and Preservation Storage Architecture

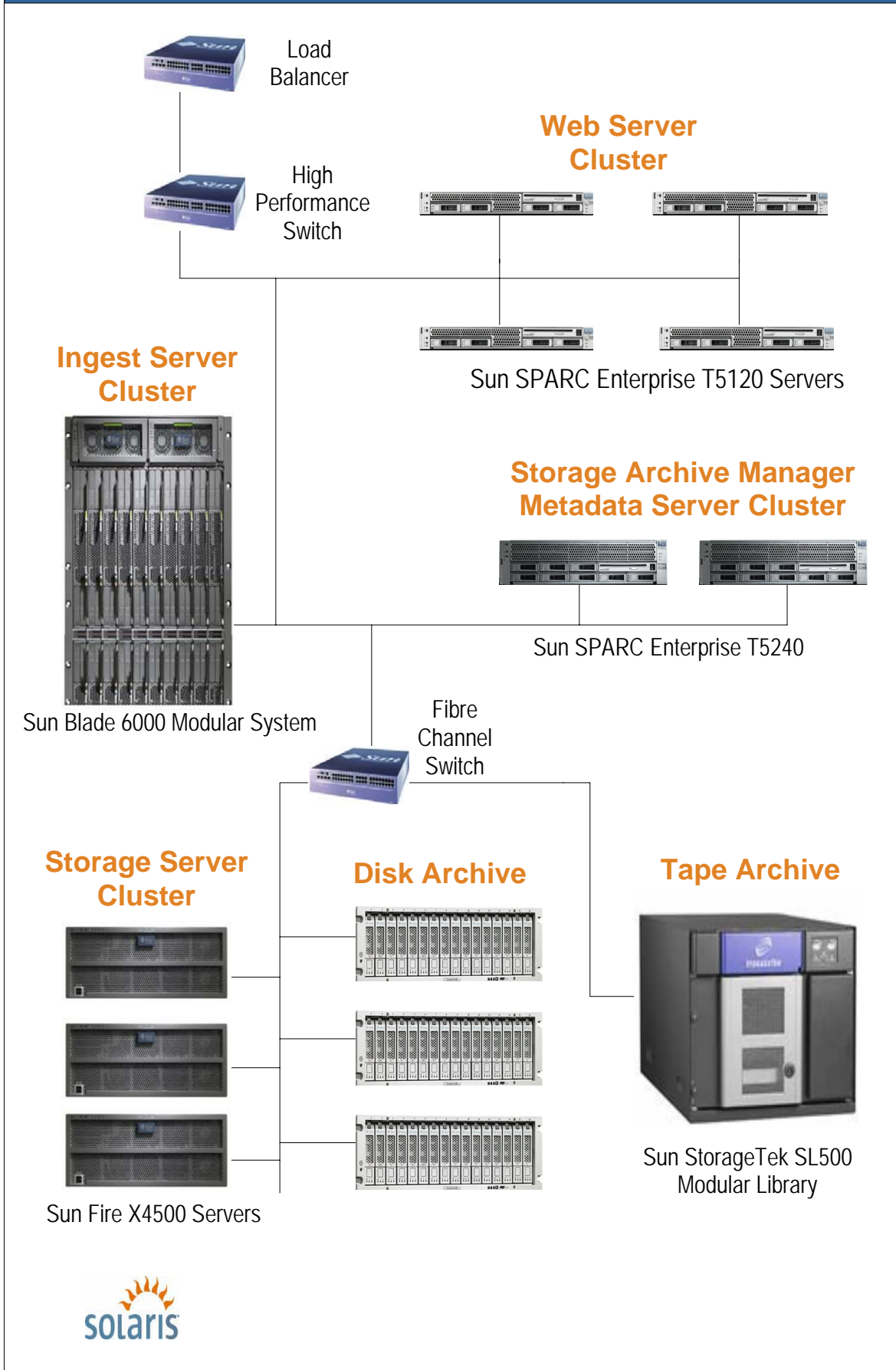


Figure 1: Digital Repository and Preservation Architecture

Implementing Preservation Services over the Storage Resource Broker

Douglas Kosovic, Jane Hunter

School of Information Technology and Electrical Engineering
The University of Queensland
douglask@itee.uq.edu.au; jane@itee.uq.edu.au

Abstract

Many international institutions and organizations responsible for managing large distributed collections of scientific, cultural and educational resources are establishing data grids that employ the San Diego Supercomputer Centre's Storage Resource Broker (SRB) for managing their collections.

Over time, maintaining access to the resources stored within SRB data grids will become increasingly difficult as file formats become obsolete. Organizations are struggling with the challenge of monitoring the formats in their SRB collections and providing suitable migration or emulation services as required. Automated methods are required that notify collections managers of objects that are at risk and that provide solutions to ensure long term access. The problem is exacerbated by the often proprietary and highly eclectic range of formats employed by scientific disciplines – many of which are too uncommon to be considered by existing national digital preservation initiatives.

This paper describes our test bed implementation of a set of preservation services (obsolescence detection, notification and migration) over a heterogeneous distributed collection of objects, stored in SRB. It also provides an evaluation of the performance and usability of the *PresSRB* system, within the context of an environmental case study.

Introduction

Existing digital preservation projects have primarily been driven by the libraries and archives communities and as such, have tended to focus on library and cultural resources – stored in repositories such as DSpace and Fedora. More recently there have been projects focusing on preservation services for resources such as CAD drawings [1] and video games [2]. Our interest is in the preservation of large scale scientific data formats increasingly being stored within the San Diego Supercomputer Centre's Storage Resource Broker (SRB) by many scientific and research organizations.

SRB is a data grid middleware system that provides a uniform interface to heterogeneous data storage resources distributed over a network. It implements a logical namespace (that points to the physical files) and maintains

metadata on data-objects (files), users, groups, resources, collections, and other items in an SRB Metadata Catalog (MCAT) which is stored in a relational database management system (e.g., PostgreSQL). Within the eScience domain, many communities are adopting SRB to implement data grids capable of managing the storage and movement of large scale collections of data and images. However, to date, no one has investigated the issues associated with maintaining long term access to digital files stored within SRB. Hence the objective of the work described in this paper is to investigate how preservation services (such as were developed within the previous PANIC [3] and AONS [8] projects) could be implemented over a collection of scientific data objects stored in SRB.

PANIC is a semi-automatic digital preservation system [3] developed at the University of Queensland that relies on semantic web services architecture. Preservation metadata associated with digital objects is generated at ingest and periodically compared with up-to-date software version, format version and recommended format registries. This enables potential object obsolescence to be detected and a notification message sent to the relevant agent. Preservation software modules (emulation and migration) were converted to web services and described semantically using an OWL-S ontology. Software agents enable the most appropriate preservation service(s) for each object to be automatically discovered, composed and invoked. The aim of PANIC was to leverage existing but disparate efforts by integrating a set of complementary tools and services including:

- Preservation metadata generation and extraction tools (e.g., JHOVE, the JSTOR/Harvard Object Validation Environment [4] and DROID [5], the National Archive's tool for performing batch identification of file formats)
- The Global Digital Format Registry (GDFR) [6]
- The UK National Archive's PRONOM project [7]

PANIC delivered a prototype system that successfully demonstrated the potential of a web services approach to automatic obsolescence detection, notification and migration. The aim of the AONS project [8] was to adapt the obsolescence detection and notification component of PANIC to generate a web service which could be applied

to multiple collection types (in particular, DSpace and Fedora) and which collections managers could easily subscribe to. AONS used preservation information about file formats and the software that these formats depend on, to determine if any of the formats within a collection are at risk. Up-to-date information about the current format and software versions was gleaned from authorized registries (PRONOM [7] and LCSDF [9]) and stored in a MySQL database. AONS then periodically checked the contents of the repository against the database to check for formats in danger of becoming obsolete. When any such formats were found, a notification report was sent to the repository manager. Because the interface between AONS and the repositories is simple, well defined and repository-independent, it was easy to deploy over different types of repositories (DSpace and Fedora).

The aim of the work described in this paper is to investigate the deployment of:

1. obsolescence detection and notification services, followed by
2. migration services

over scientific data/objects stored within SRB. This work will be carried out through the development of the *PresSRB* prototype and its evaluation through an environmental case study.

The use of SRB also raises a number of new and challenging issues that need to be considered, including:

- The storage of preservation metadata within MCAT;
- Recommended format registries for data associated with specific scientific disciplines;
- Obsolescence detection and migration services for multiple versions of the same object, stored at distributed locations within a SRB data grid.

An Environmental Case Study

Remote sensing satellite images are typical of many scientific data sets. They are represented in a wide range of formats – both open and proprietary, depending on the organization or satellite operator producing the images. Formats include: CCRS, EOSAT, HDF (Hierarchical Data Format), Fast-L7A, CEOS, ERDAS Imagine and GeoTIFF (Geographic Tagged Image File Format). GeoTIFF is the most popular standardized file format for GIS applications – suitable for storage and transfer across operating system environments and applications. It is open, public domain and non-proprietary. GeoTIFF embeds georeferencing information (e.g. projection, datums and ellipsoids, coordinate values) as metadata within the TIFF (Tagged Image File Format) file [10]. As the GeoTIFF format is fully compliant with the TIFF 6.0 specification, applications that don't know about the GeoTIFF tags will be able to open them like any other TIFF file.

For evaluating the PresSRB system, we decided to work with existing users of SRB – the Centre for Remote

Sensing and Spatial Information Science (CRSSIS) at the University of Queensland. CRSSIS are using SRB for the storage and analysis of Landsat 5 satellite images provided by the Queensland Department of Natural Resources and Water in the ERDAS Imagine (.img) file format.

The Landsat 5 satellite has an onboard sensor called the Thematic Mapper (TM). The TM sensor records the surface reflectance of electromagnetic (EM) radiation from the sun in seven discrete bands. Reflectance is the ratio of outgoing light reflected from the land surface to the incoming light from the sun. Mosaics of Landsat 5 satellite image data are provided as 6 layer ERDAS Imagine files. The various layers and corresponding wavelengths are shown in table 1.

Image layer	Landsat Band	Wavelength (µm)
1	1	0.45 - 0.52 Blue
2	2	0.52 - 0.60 Green
3	3	0.63 - 0.69 Red
4	4	0.76 - 0.90 near infrared
5	5	1.55 - 1.75 shortwave infrared
6	7	2.08 - 2.35 shortwave infrared

Table 1: Landsat TM ERDAS Imagine Layers

ERDAS Imagine is a commercial raster image processing and remote sensing geographic information system (GIS) application. The current version is 9.2 and is available for Microsoft Windows. In the past, both Unix/X-Windows versions were also available. For images that require more than 4GB of disk space, Imagine creates two files: the .img file contains the traditional superstructure, but the actual raster data is kept in a separate file which has an extension .ige [11]. The ERDAS Imagine file samples that were used in this case study ranged from 6.6 GB (28189 x 38828 pixels) to 20 GB (45144 x 72111 pixels).

The problem with ERDAS Imagine files are that they are very large, depend on proprietary software and are difficult to manage. To maximize long term access and availability, they should ideally be converted to GeoTIFF (both full-size and thumbnail for previewing).

Although the TIFF specification allows for multi-spectral imagery (more than 3 bands), many software applications are unable to handle multi-spectral TIFF files. To overcome this limitation, many satellite image providers deliver two GeoTIFF files, one with red, green and blue bands and another with near infrared, red and green bands [12]. For the PresSRB prototype, we present the collections manager with the option of either selecting 3 bands (RGB) or 4 bands (RGBA) for converted GeoTIFF files generated by the migration service. In addition, because TIFF 6.0 uses 32bit unsigned offsets, it is limited to a maximum files size of 4GB. Our SRB data grid contained a number of ingested GeoTIFF files that were bigger than 4 GB. For these files we used BigTIFF (aka TIFF-64) – a proposed standard for TIFF data bigger than 4GB in file size. There

are now a growing number of geospatial tools and libraries which are able to support it.

For the conversion from Imagine to GeoTIFF format, we used GDAL (Geospatial Data Abstraction Library). GDAL is a translator library for raster geospatial data formats that is released under an X/MIT style Open Source license by the Open Source Geospatial Foundation [13].

GeoTIFF format is registered in PRONOM as a distinct format, but LCSDF has no format description for GeoTIFF. The ERDAS Imagine formats are not registered in the PRONOM, LCSDF or GDFR registries. Ideally there should be a recommended format registry for geospatial data, that provides best practice guidelines for the archival and curation of geo-spatial data.

System Architecture

Figure 1 illustrates the overall architecture of the PresSRB prototype and the various software layers that interact with the underlying SRB Data Grid.

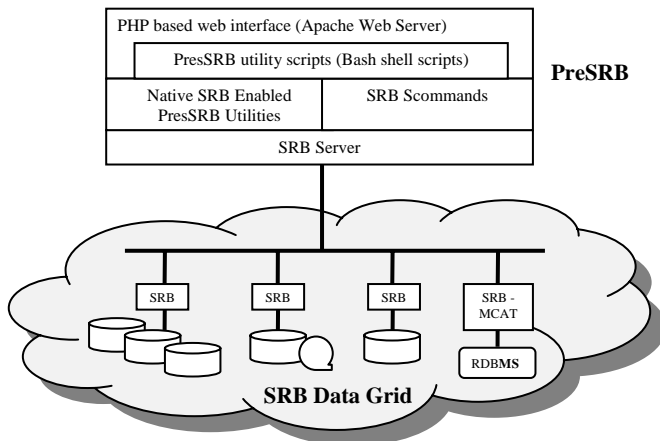


Figure 1: PresSRB System architecture

The PresSRB prototype was implemented on a PC running RedHat Enterprise Linux (RHEL) 4. Significant effort is required to setup and deploy SRB, so to simplify the deployment of SRB on other RHEL systems, SRB-3.4.2 was packaged into a number of RedHat Package Manager (RPM) files.

SRB Data Grid

The SRB is a data grid middleware system that provides a uniform interface to heterogeneous data storage resources distributed over a network. It implements a logical namespace (that points to the physical files) and maintains metadata on data-objects (files), users, groups, resources, collections, and other items in an SRB Metadata Catalog (MCAT) stored in a relational database management system.

SRB Scommands

Scommands are command-line SRB client utilities for accessing SRB data and metadata. Most *Scommand* names have a “S” prefix. *Scommands* are the most powerful and flexible of the SRB clients that come with the SRB source code. They are ideal for batch jobs, scripting and PHP wrappers.

Shell scripts are extensively used in the PresSRB prototype to wrap Scommands and the previously mentioned SRB-enable utilities, in order to provide much of PresSRB’s functionality and to perform batch operations.

SRB-enabled Utilities

The simplest approach to providing SRB support to a non-SRB application is to copy or replicate data from SRB space to a local file space, and then provide the application with the local filename. This approach works very well when an entire file is to be processed, but if, for example, only the first 26 bytes of a 20GB ERDAS Imagine .ige file is required for format identification, it’s very wasteful both with respect to bandwidth and performance.

For the PresSRB prototype, the Linux **file** command was used to identify file formats [14]. It was modified to make SRB client library calls instead of Unix file I/O calls. The benefit is that the stock **file** command reads at most 256kB of data to identify a file - so too does the modified **file** SRB-enable command hereafter referred to as **Sfile**.

Similarly the GDAL **gdalinfo** georeferencing meta-data extractor used on ERDAS Imagine and GeoTIFF files was modified to add native SRB support, as it also only needs to read a small portion of the files.

Each ERDAS Imagine file was migrated to two new files: a low resolution, 1% sized preview GeoTIFF file and a equivalent resolution GeoTIFF file.

For file format conversion in PresSRB, an Scommand-based shell script wrapper was initially used which retrieved a file from SRB space using the **Sget** command, then invoked the **gdal_translate** utility to perform the conversion on a local file and then upload a converted file to SRB space using **Sput**. For transferring large files using **Sput** and **Sget**, it was possible to take advantage of SRB’s parallel I/O capabilities (multiple threads each sending a data stream over the network) which made SRB significantly faster than HTTP, FTP, SCP or even NFS.

We also tested adding native SRB support to the **gdal_translate** utility. This was beneficial for the conversion of large ERDAS Imagine files to BigTIFF/GeoTIFF file of the same resolution. This was because only 3 of the 6 bands of the ERDAS Imagine files are processed in the conversion.

Generation of small GeoTIFF preview files (scaled down to 1% in both the horizontal and vertical resolution from the original ERIDAS Imagine files) also had a significant benefit, since only a small “overview” needs to be read which is only a tiny fraction of the ERIDAS Imagine file size. For some of the sample ERIDAS Imagine files we used, they had up to nine overviews ranging from 34x84 to 8504x21269 pixels in size.

PHP Web interface

The SRB-3.4.2 source code comes with a sample ScmdWrapper PHP class which enables wrapping SRB Scmmands. The SRB authentication, browsing, searching and ingestion components of the PresSRB Web interface employs the ScmdWrapper PHP class. A number of the PresSRB Scmmand-based shell scripts which are invoked by a scheduler component are also able to be directly executed from the PresSRB Web GUI. This approach improves system performance (over the Tomcat-based Java approach used by AONS) because it requires less memory and overheads when dealing with large files.

System Implementation

Figure 2 below illustrates the main components of the PresSRB system. The six main components (described in the next 6 subsections) are:

1. Format identification and Preservation metadata
2. Format, Software and Recommendation Registries
3. Obsolescence Detection
4. Migration and Preview
5. Scheduler
6. Web GUI

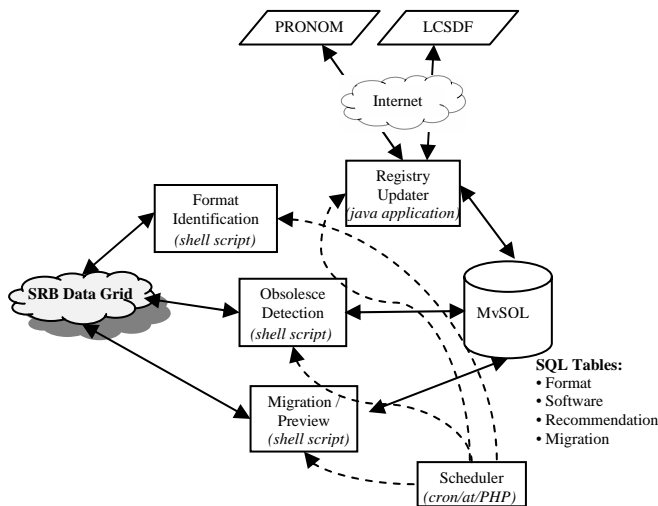


Figure 2: PresSRB non-GUI components

Format Identification and Preservation Metadata

Format identification in the PresSRB prototype involves identifying the file format of SRB data objects (i.e. files in SRB space), extracting georeferencing metadata if

available and then populating the SRB data objects with PresSRB specific SRB user defined metadata. This is all performed by a shell script which invokes the following three commands:

- SRB enabled **file** command (**Sfile**) for file identification.
- SRB enabled **gdalinfo** for georeferencing metadata extraction.
- **Sufmeta** Scmmand which provides the facility for inserting, deleting, updating SRB user-defined metadata (attribute name-value-unit triplets).

The **Sfile** utility is unaware of BigTIFF files and ERIDAS Imagine .img and .ige files. In order to recognize these formats, the `magic/Magdir/images` file from the `file` source code needed to be modified to add signature recognition for these formats and then the file source code recompiled to generate a magic file with the new signatures.

The **Sfile** command currently recognizes GeoTIFF files as just generic TIFF files. Instead of modifying the **Sfile** command to specifically recognize GeoTIFF files, the shell script wrapper uses the **gdalinfo** command to confirm if they are GeoTIFF files.

The shell script wrapper maps the output of the **Sfile** command on a SRB data object to a PRONOM Persistent Unique Identifier (PUID) for a set of known file formats. If the shell script is able to determine the PUID of a SRB data object, it then inserts a SRB user defined attribute called “pressrb_PUID” on that data object with the value set to the PUID. If the **Sfile** command is able to determine the mime type of a data object, then a “pressrb_mime_type” is also inserted.

To avoid name space collisions of SRB metadata attribute names which might already be used in an existing SRB data grid, the attribute names are prefixed with “pressrb_”.

As PUIDs for BigTIFF and ERIDAS Imagine file formats haven’t been assigned yet in the PRONOM repository, we are currently using a local format repository MySQL table to define the interim PUIDs listed in Table 2. We intend submitting a request to the managers of PRONOM to have PUIDs assigned for these formats.

PUID	Format Name	Extension
x-fmt/10000	BigTIFF	tif, tiff
x-fmt/10001	BigTIFF/GeoTIFF	tif, tiff
x-fmt/10002	Erdas Imagine	img
x-fmt/10003	Erdas Imagine - Large Raster Spill File	ige

Table 2: Local File Format Repository

For example, given the following ERIDAS Imagine SRB data object file:

```
srb:/home/srb.demo/sample.img
```

The format identification shell script wrapper will invoke the **Sufmeta** command similar to command-line that follows, which sets the “pressrb_PUID” attribute to the appropriate PUID value:

```
Sufmeta -d pressrb_PUID "x-fmt/10002" sample.img
```

As we are dealing with very large raster images, the output target file can easily exceed the maximum file size permitted for that format (e.g., a TIFF file cannot exceed 4GB – if it exceeds this size, it should be represented as a BigTIFF file). Metadata that can predict the file size for converted files is very important for the migration service.

Table 3 shows sample raster image metadata for an ERDAS Imagine file (name, value, units triplets). This metadata was extracted using **gdalinfo**. This file is to be migrated to a GeoTIFF image (34014 pixels x 85075 lines x 3 bands/bytes). The estimated size of the output file will be larger than 4GB. In this situation, a warning message is displayed which includes the estimated number of files of excess size.

Name	Value	Units
x_resolution	34014	pixel
y_resolution	85075	pixel
num_bands	6	

Table 3: General Raster Image Metadata

Georeferencing metadata is also extracted from ERDAS Imagine files using the **gdalinfo** command. Table 4 lists the georeferencing metadata for a sample SRB data object that is stored in MCAT. This georeferencing metadata is not currently being used in PresSRB as preservation metadata. But it may be required in the future to prevent lossiness - if we convert to file formats which only supply limited georeferencing metadata.

Name	Value	Units
proj_coords	Transverse Mercator	
latitude_of_origin	0	Degrees
central_meridian	141	Degrees
scale_factor	0.99959999	
false_easting	500000	M
false_northing	10000000	M
pixel_size_x	25.0	m/pixel
pixel_size_y	-25.0	m/pixel
x_axis_rotation	0	Degrees
y_axis_rotation	0	Degrees
easting	178412.5	M
northing	8903562.5	M

Table 4: Georeferencing Metadata

Format, Software and Recommendation Registries

PresSRB re-uses the AONS I registry Java code with some minor modifications. This registry interface performs the role of retrieving:

- the format and software information from PRONOM registry; and
- format, software and recommended format information from the LCSDF registry.

The AONS I registry code is a web crawler which retrieves data from the PRONOM and LCSDF web sites via HTTP and transforms it into XML which is then inserted into the format registry, software registry and recommendation registry (MySQL tables) [8].

Because the BigTIFF and ERDAS Imagine formats are not represented in the PRONOM and LCSDF registries, and GeoTIFF is not represented in the LCSDF registry, MySQL tables were created to supplement the external registries for these formats. For example, the local format registry table is populated with data based on the values shown in Table 2.

Obsolescence Detection

Because PresSRB re-uses the AONS I format registry, software registry and recommendation registry, the four types of obsolescence warnings generated by AONS I are also generated within PresSRB. These are:

- Format has a new version
- Format not supported by any software
- Format is proprietary
- Format supported by obsolete software.

Figure 3 illustrates an example report warning of proprietary format obsolescence.

PresSRB Obsolescence Report				
SRB Collection: /A/home/srb.demo/2002/				
Format is proprietary				
PUID (PRONOM Unique Identifier)	Name	Total affected items	Number of unmigrated files	Recommendation
x-fmt/10002	Eradas Imagine	6	6	Migrate file(s) to non-proprietary GeoTIFF file(s)

Figure 3: Sample PresSRB Obsolescence Report

In the future, PresSRB will also generate the following warning:

- Format has no encapsulated georeferencing metadata

This situation would occur when a raster image file which has no embedded georeferencing metadata is co-located in an SRB collection with a separate file which contains the metadata.

Input Format	Output Format	Quality Ranking	GDAL Input Driver	GDAL Output Driver	Description
ERDAS .img	GeoTIFF	10	ERDAS	TIFF	If RGBA photometric interpretation option is selected for the converted files, the files may not be able to be opened or handled correctly by many software applications
ERDAS .img	JPEG2000	9	ERDAS	JP2KAK	This converter requires GDAL to be built with the Kakadu SDK which can be purchased from http://www.kakadusoftware.com/
ERDAS .img	JPEG2000	5	ERDAS	JPEG2000	This converter does not support generating output files that are > 2GB. Consider using the commercial Kakadu (JP2KAK) GDAL driver instead if JPEG2000 is required.

Table 5: Migration table (PUIDs replaced with actual format name for clarity)

Migration and Preview

The migration service converts raster image files from one format (ERDAS Imagine files) to another (GeoTIFF), preserving the resolution of the original image file, although not necessarily preserving the number of bands of the original.

The preview service provides a scaled down GeoTIFF version of the original geospatial raster files - the intention being to provide small preview files which are only a few MB in size as opposed to GBs in size

Unlike file systems which do not allow a folder and a file to have the same path name, SRB does allow data objects to have the same logical path name as SRB collections. So for example, given the following ERDAS Imagine SRB data object:

```
/home/srb.demo/sample.img
```

SRB allows us to generate the following migrated and preview data objects located in a sample.img data collection, while leaving the original data object intact:

```
/home/srb.demo/sample.img/geotiff_rgb.tif
/home/srb.demo/sample.img/preview_geotiff.tif
```

Currently the migration and preview services are limited to the input and output formats supported by GDAL [15]. Furthermore, they are limited by the drivers GDAL was built against and the features that this driver supports. For example, the JPEG2000 driver based on the free JasPer JPEG2000 library has a maximum file size support of 2GB, while the JP2KAK JPEG2000 driver which requires the commercial Kakadu library has an unlimited file size.

The migration service consists of two main components: a discovery component and a conversion provider component.

The intention of the migration discovery component is to present the collection manager with migration options for a specified file format. This is achieved by querying the migration MySQL table (similar to what is shown in Table

5.) which lists the various GDAL conversions and provides a description of the quality of the conversion. It contains a quality ranking field with a possible value between 0 and 10 which used to rank the available conversion providers presented to the collection manager.

The output formats supported by GDAL utility is listed using the `--formats` switch. For example the JP2KAK JPEG2000 conversion provider will be listed, but it won't be a selectable option. However information on where to purchase the Kakadu JPEG2000 library is presented.

If there is a reduction in the number of bands from the original to the migrated file (e.g., 6 bands in ERDAS Imagine and 3 bands in GeoTIFF), then this information is also presented to the collection manager.

If any of the generated GeoTIFF files are > 4GB, warnings are displayed that warn that some of the files will need to be converted to BigTIFF.

Scheduler

The PresSRB services were designed to be either executed directly from the command-line or via a thin PHP-based web wrapper. Because they can be invoked via the command line, the standard scheduling services available on unix-like operating systems can also be used.

For scheduled operations that are required to be executed periodically in some sort of recurring pattern, the **cron** scheduler is used. For operations which are required to be performed once at some time in the future, the **at** command is used. Once a scheduled job is completed, a email summary is sent to the collection manager.

Web GUI

The PresSRB GUI consists of a number of components which are described below.

Authentication. An authentication Web page is used to authenticate content managers Only the authenticated content managers have the right to generate the

obsolescence reports and execute migration services. Other users are only permitted to browse, query and view the collections and reports.

Browsing. A simple Web interface for browsing through a SRB collection is provided. When a content manager selects a collection or data object, they also have access to menu items that allow them to invoke and schedule the obsolescence and migration/preview services.

Querying. The standard SRB user-defined metadata queries can be performed via a simple web interface which returns a list of SRB Objects that match the query.

Ingestion. The PresSRB prototype provides a web form to enable ingestions of individual files into SRB space. File upload uses the multipart POST method. A PHP script receives the upload file and then ingests it into SRB space using the **Sput** Scommand. For PHP, the default maximum file size value is only 2MB. Thi would need to be increased for the handling of larger files. Although this can be set as high as 2GB, it's much more efficient to use a dedicated SRB client or the *Scommands* to ingest very large files. Most web-administrators would also discourage setting this value too high.

Obsolescence The obsolescence web page displays the last generated obsolescence report. Generating new reports is handled by the scheduler (described below). Figure 5 shows the scheduler page for the obsolescence detection scheduler. Users are able to specify the frequency of job execution via the GUI.

Migration and Preview. The dynamically generated Migration Web page is shown in Figure 4. For a given collection, this page displays the current file formats requiring migration (and the number of files) and provides a pull-down menu of migration services that can be applied. When converting to GeoTIFF format, users can specify the number of bands (RGB or RGBA). It also lists the recommendations from the registries (if available) and any issues associated with the migration service. Selecting the Schedule button, will schedule the underlying migration script.

Scheduler. The Scheduler enables users to schedule either obsolescence detection or migration scripts. A crontab configuration file specifies how to execute commands on a particular schedule and the addition, modification and removal of jobs. A PHP wrapper for the crontab command is used within the PresSRB Web GUI; so that users won't need to know the intricacies of the crontab file syntax. Figure 5 below illustrates the scheduler Web user interface.

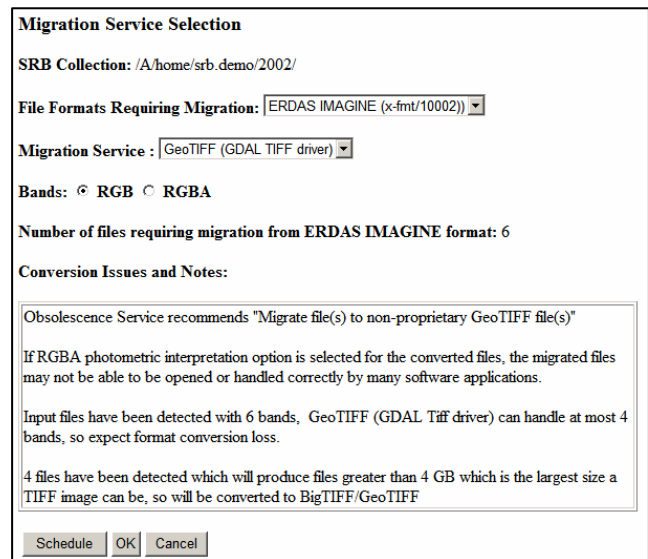


Figure 4: Migration Service Selection GUI

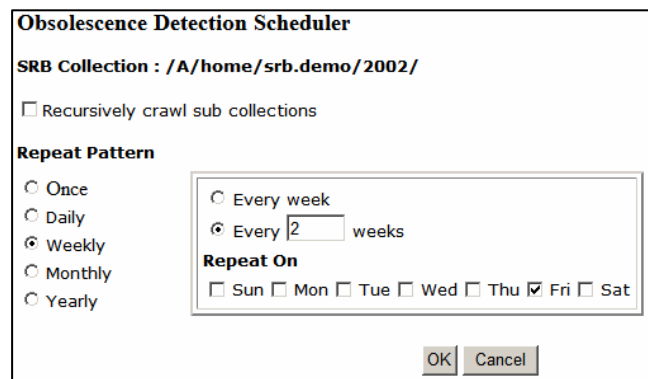


Figure 5: PresSRB Scheduler GUI Interface

Discussion and Conclusions

One of the greatest challenges with regard to scientific data is the lack of preservation services or support for many scientific data formats. There is a need for an initiative to establish a registry of recommended formats for scientific data within different disciplines. More specifically, there are currently no registries which provide recommendations for GIS based file formats. For example, neither PRONOM nor LCSDF support ERDAS Imagine files. Within PRONOM, GeoTIFF is just a placeholder and no real information is provided. In LCSDF, GeoTIFF isn't registered, even as a sub-format of TIFF.

Also particularly within the scientific domain, we are entering an era in which many files formats are reaching their file size limits (either 2 or 4 GB). In these cases, the current file format can no longer be used as a recommended file format for migration. Additionally, for many scientific datasets, the generic preservation metadata is inadequate and needs to include specialised metadata, e.g., for raster images, the resolution of the original file is

important. Georeferencing metadata is also significant but not currently supported in standardized preservation metadata schemas.

To conclude, SRB is an ideal infrastructure for dealing with large files, especially because it expedites the copying and replicating of large scale files using parallel data transfer. It provides an ideal infrastructure for preservation based on the LOCKSS [16] approach. However it is lacking in support for the preservation services required to ensure long-term access to many of the file formats being stored within SRB. Within the PresSRB prototype described in this paper, we have implemented and evaluated obsolescence detection and notification services and migration services for a particular environmental case study. We have demonstrated how this can be achieved by integrating external services using wrappers around the SRB Scocommands and by adding native SRB support to existing utilities. Significant further work is required to evaluate and implement similar approaches for data archived within SRB by other scientific disciplines such as astronomy, biology and earth sciences.

References

1. Ball, A. and M. Patel, *Approaches to Information Curation in Engineering*, in *Knowledge and Information Management Through-Life*. June 2008: Institution of Mechanical Engineers, London.
2. Guttenbrunner, M., C. Becker, and A. Rauber, *Evaluating Strategies for the Preservation of Console Video Games*, in *iPRES2008*. 2008: London, UK.
3. Hunter, J. and S. Choudhury, *A Semi-Automated Digital Preservation System based on Semantic Web Services*. Joint Conference on Digital Libraries, JCDL, 2004: p. 269-278.
4. *JHOVE - JSTOR/Harvard Object Validation Environment*. [cited; Available from: <http://hul.harvard.edu/jhove/>].
5. The National Archives of the UK. *DROID Digital Record Object Identification*. [cited; Available from: <http://droid.sourceforge.net/wiki/>].
6. *Global Digital Format Registry*. [cited; Available from: <http://www.gdfr.info/>].
7. *PRONOM*. [cited; Available from: <http://www.nationalarchives.gov.uk/pronom/>].
8. Curtis, J., et al., *AONS - An Obsolescence Detection and Notification Service for Web Archives and Digital Repositories*. Special issue on Web Archiving for the New Review on Hypermedia and Multimedia (JNRHM), January 2007. **13**(1): p. 39-53.
9. *LCSDF (Library of Congress Sustainability of Digital Formats)*. [cited; Available from: <http://www.digitalpreservation.gov/formats/>].
10. Ruth, M. *GeoTIFF FAQ Version 2.3*. 2005 [cited; Available from: <http://www.remotesensing.org/geotiff/faq.html>].
11. *Erdas Imagine .ige (Large Raster Spill File) Format*. [cited; Available from: http://home.gdal.org/projects/imagine/ige_format.html].
12. Beaty, P. *What is wrong with my GeoTIFF?* 2008 [cited; Available from: <http://field-guide.blogspot.com/2008/05/what-is-wrong-with-my-geotiff.html>].
13. *GDAL - Geospatial Data Abstraction Library*. [cited; Available from: <http://www.gdal.org/>].
14. *Fine Free File Command*. [cited; Available from: <http://www.darwinsys.com/file/>].
15. *GDAL Raster Formats*. [cited; Available from: http://www.gdal.org/formats_list.html].
16. *LOCKSS (Lots of Copies Keep Stuff Safe)*. [cited; Available from: <http://www.lockss.org/>].

Embedding Legacy Environments into A Grid-Based Preservation Infrastructure

Claus-Peter Klas, Holger Brocks, Lars Müller, Matthias Hemmje

FernUniversität in Hagen

Universitätsstrasse 1

58097 Hagen, Germany

{Claus-Peter.Klas, Holger.Brocks, [Lars.Müller](mailto:Lars.Mueller), Matthias.Hemmje}@FernUni-Hagen.de

Abstract

The SHAMAN project targets a framework integrating advances in the data grid, digital library, and persistent archival communities in order to archive a long-term preservation environment. Within the project we identified several challenges for digital preservation in the area of memory institutions, where already existing systems start to struggle with e.g. complex or many small objects. In order to overcome these, we propose a grid based framework for digital preservation. In this paper we describe the main objectives of the project SHAMAN and the identified challenges for such a heterogeneous and distributed environment. We on the one hand assess in a bottom-up approach the capabilities and interfaces of legacy systems and on the other hand derive requirements based on the project's objectives. Our investigation is focused to the integration of storage infrastructures and distributed data management. In the end we derive a service-oriented architecture with a grid-based integration layer as an initial approach to manage the challenges.

The SHAMAN Project

As part of the European Commission's 7th Framework Program, the SHAMAN (Sustaining Heritage Access through Multivalent ArchiviNg) project targets a framework integrating advances in the data grid, digital library, and persistent archival communities in order to attain a long-term preservation environment which may be used to manage the ingest, storage, preservation, access, presentation, and reuse of potentially any digital object over time. Based on this framework, the project will provide application-oriented solutions across a range of sectors, including those of digital libraries and archives, design and engineering, as well as scientific data and information management.

The SHAMAN project will integrate the automated handling of technology evolution with data analysis and representation mechanisms in a way which will uniquely enable multiple user communities to preserve and reuse data objects, in whatever format, which are deposited in the preservation environment.

The project will furthermore provide a vision and rationale to support a comprehensive *Theory of Preservation* that may be utilized to store and access potentially any type of data, based on the integration of digital library, persistent archive, knowledge representation, and data management technologies. In addition SHAMAN will supply an infrastructure that will provide expertise and support for users requiring the preservation and re-use of data over long-term periods of time. Within this infrastructure the project will also develop and implement a grid-based production system that will support the virtualization of data and services across scientific, design and engineering, document, and media domains. Finally three Integration and Demonstration Subprojects (ISPs) supporting the *Memory Institutions* (ISP-1), *Design and Engineering* (ISP-2) and *eScience* (ISP-2) domains are used to analyze their ecology of functional (and non-functional) requirements and to identify a core set of relevant digital preservation usage scenarios. These ISPs foster the systematic integration and evolution of project results towards the targeted SHAMAN framework and its prototypical application solutions, i.e. they drive the horizontal integration of RTD contributions.

In this paper, we will focus on ISP-1 Document Production, Archival, Access and Reuse in the Context of Memory Institutions for Scientific Publications and Governmental Document Collections, which trials and validates the SHAMAN approach along the business purposes of scientific publishing, libraries, and parliamentary archives.

We will present challenges which are derived from the preliminary results of the top-down requirement analyzes of ISP-1. From the bottom-up technology perspective we have conducted an initial assessment of the capabilities and interfaces of the systems employed inside and outside the SHAMAN consortium which hold relevant digital collections, but also solutions supporting access (i.e. searching and browsing), resource discovery and collection management. We will then elaborate on the specific technological challenges of integrating heterogeneous storage infrastructures and distributed data management and present a first conceptual integration-approach based on a grid-based integration layer and service-oriented architectures for resolving these issues.

Integration Requirements

The goal of this paper is to describe digital preservation legacy technology and existing application solutions as well as a draft integration concept for embedding such legacy environments into an overall preservation infrastructure like the SHAMAN framework. To evaluate this integration concept we need to provide an assessment scheme which represents general digital preservation requirements, but also specific challenges derived from integration of individual, complex systems and processes within the SHAMAN context. The following generic integration requirements represent overall conceptual goals or success criteria for the SHAMAN framework, which are refined and complemented by more specific challenges from the ISPs:

- **Integrity** - The main goal of preservation environments is to maintain the persistence of digital objects. Integrity refers to maintaining their completeness and immutability. A preservation environment has to provide adequate measures for maintaining the integrity of its digital objects.
- **Authenticity** - Authenticity corresponds to the genuineness of an object. An object is considered as genuine if certain properties can be attested which confirm its identity. A preservation environment must prevent unauthorized manipulation of its objects in order to guarantee their authenticity.
- **Search & Browse** - Besides safe-keeping its digital objects, a preservation environment also needs to provide access to its collections. This requires persistent identifiers and sophisticated search methods to find and access particular objects.
- **Interpretability** - Technological advancements leads to the aging of digital object formats. The careful selection of allowed formats according to various criteria enables the long-term interpretability of the content of digital objects. Furthermore, preservation environments need to support strategies for dealing with technological obsolescence.

- **Virtualization** - The integration of distributed information systems requires coherent management of the heterogeneous systems and collections. A federated preservation environment needs to abstract from the idiosyncrasies of its constituting peers, while maintaining full control over processes and objects, including their significant properties.

Following these general requirements the next section describes the specific scenarios for memory institution in ISP-1.

ISP1 - Memory Institutions

Within SHAMAN's ISP-1 scenario we need to provide long term preservation for three memory institutions (2 libraries, 1 archive), the German National Library (DNB), the Niedersächsische Staats- und Universitätsbibliothek Göttingen (SUB) and the Flemish Parliament (FP) governmental archive. All of them are running existing individual solutions. As a grid-based data management system iRODS will be assessed and trialed as the successor of earlier SRB technologies, which already is used in Europe and the US for very large file repositories. We will evaluate its appropriateness for virtualizing the storage layer of the SHAMAN preservation framework architecture, also with respect to its capabilities for integrating existing legacy systems with proprietary data schemas. The initial goal of the iRODS assessment, is to embed the existing repositories and archival systems from the DNB, SUB and FP as an active node within the iRODS data grid. In the following sections we will describe the legacy systems and discuss possible solutions based on existing tools and grid systems.

Existing Storage and Long-Term Preservation Systems

The SHAMAN application scenarios require the integration of various types of existing and upcoming systems. Examples of these systems are institutional repositories like Fedora and DSpace, the KOPAL long-term digital information archive, standard database storage systems as well as access support systems such as DAFFODIL or Cheshire. These systems have to be assessed individually, but also the resulting composite infrastructures have to be evaluated according to the challenges described above. We will discuss the above named institutional repositories, archive systems and access systems closer in the next sections. The grid-based systems under evaluation follow the legacy system description.

Institutional Repositories

Institutional Repositories are used for managing documents and collections within scholarly environments, such as universities and libraries. As production systems they need

to be integrated in a transparent way, without impeding or compromising their primary functions.

Current global players on institutional repositories are Fedora and DSpace.

Fedora

Fedora (Flexible Extensible Digital Object and Repository Architecture) represents a repository enabling archival, retrieval and administration of digital objects and their metadata via web services. It is developed at the Cornell University and the University of Virginia. Within Fedora a digital object is a container for different components. These are a unique identifier, descriptive metadata, data streams, and disseminators. Each container consists of at least one data stream including metadata in Dublin Core format. A data stream can also be a URL. An object can also contain disseminators connected to a data stream to generate dynamically different views, e.g. a black/white picture of a color picture.

Fedora also supports integrity via checksums and authenticity of digital objects. Redundancy is based on replication on a second Fedora system. Archiving, retrieval, and administration is based on SOAP and REST web services. To access the metadata an OAI-PMH server is integrated to provide access to other systems. The system is OAI-PMH conform and supports ingestion of SIPs (digital objects with METS data) objects.

There are currently 127 Fedora projects and 25.000+ downloads have been counted last year according to the Fedora Wiki.

DSpace

DSpace is like Fedora an institutional repository developed by Hewlett-Packard and the Massachusetts Institute of Technologies as open source project. Objects (items) are stored in collections structured by communities and sub communities.

Each item represents an archived object, including metadata and further files like thumbnails of the original picture. Here also checksums are used to check the integrity of stored objects. Metadata is supported via Dublin Core and other formats can be transformed. An OAI-PMH supports access of metadata, so DSpace can be used as data provider. Objects can be stored in the local file system or via SRB / iRODS data grid technology.

Search and browse functionality is provided by a web interface and DSpace uses persistent identifiers.

Currently DSpace exists in 324 installations in 54 countries with approx. 2.561.082 Documents according to the DSpace Wiki.

Archival Systems

Long-term archival systems are complex IT systems with idiosyncratic processes and information structures. With their ability to provide bit-stream preservation functions at various service levels, archival systems will be embedded as specialized storage nodes which offer higher levels of data security.

A running long-term archival system is operated at the German national library, called KOPAL. As central archival library and national bibliographic center for the Federal Republic of Germany the German National Library DNB has to collect and archive also all electronic publications appearing in Germany since 2006. To comply with this assignment the DNB builds up in co-operation with other national and international memory institutions an IT-infrastructure for archiving and long-term preservation of digital objects. In its current state this infrastructure consists of a repository system for collecting digital objects, bibliographically preparing them and allowing access for external users. All objects are then archived in a back-end archival system for long-term preservation.

This long-term archival system was developed cooperatively with the Niedersächsische Staats- und Universitätsbibliothek Göttingen, the Gesellschaft für Wissenschaftliche Datenverarbeitung mbH Göttingen (GWDG) and IBM Germany within the project KOPAL (2004 - 2007). The technical realization is based on prior work accomplished since 2000 in a joint development project of the Koninklijke Bibliotheek (Royal Dutch Library) and IBM.

The core component of the system is the DIAS archive developed by IBM. DIAS implements core components of the Reference Model for an Open Archival Information System OAI. Hosted at the GWDG at Göttingen this multi-client capable system provides independent access for each partner from any location via well defined interfaces. DIAS itself consists of standard applications by IBM, DB2, WebSphere, and the Tivoli Storage Manager.

The project partners of KOPAL implemented supplementary open source software on top of the DIAS-Core, the so-called kopal Library for Retrieval and Ingest (koLibRI). Realized in Java KoLibRI provides tools for automating archiving tasks like ingest and access of digital objects in a flexible and modular manner.

Currently the KOPAL archival system is transferred into the productive use by the German National Library. It will be integral part of a more complex repository and archival system which cover the whole process from data collection via data preparation, data archiving, data access to data presentation. This repository system itself is integrated in the library system with its several tasks and services.

For the communication with the outside world there are several interfaces provided or in preparation including web forms, SRU and services based on OAI (Open Archival Initiative), especially the OAI-PMH protocol. With these interfaces the foundations are complied to integrate the DNB repository system into subordinated infrastructure networks as it is planned in the integrated project SHAMAN and to provide and exchange data and metadata within these networks.

The Flemish Parliament document storage consists of several classical databases, which can be searched via a web interface. We have to investigate their preservation proprietary solution.

Access Support Systems

Content-based access support represents a fundamental requirement for SHAMAN, in addition to traditional metadata-based search and browsing functions. The main challenge within a federated environment is to keep the retrieval index consistent and up-to-date.

Cheshire

Within SHAMAN we plan to integrate Cheshire, a full-text information retrieval system based on a fast XML search engine. On the basis of indexes it provides access to the essential search and browse functionality of digital libraries. Cheshire's development started 10 years ago at these UC Berkley and currently is run in version 3 by the University of Liverpool. It supports several protocols like Z39.50, SRW/SRU or OAI-PMH for access of metadata.

DAFFODIL

To provide users with the ability to find and access their preserved information we will utilize the DAFFODIL system .

DAFFODIL is a virtual digital library system targeted at strategic support of users during the information seeking and retrieval process (see [Fuhr et al.2002] and [Klas2007]). It provides basic and high-level search functions for exploring and managing digital library objects including metadata annotations over a federation of heterogeneous digital libraries. For structuring the functionality, we employ the concept of high-level search activities for strategic support and in this way provide functionality beyond today's digital libraries. A comprehensive evaluation revealed that the system supported most of the information seeking and retrieval aspects needed for scientists' daily work. It provides a feature rich and user-friendly Java swing interface to give access to all functionalities. Furthermore a Web 2.0 browser interface enables the main but not all functions of the Java interface for easy access. Besides the main functionality of federated search and browse in distributed and heterogeneous data sources and a personal library, further functionalities like browsing co-

author networks, thesauri, conference & journal browser and collaborative functions are already implemented and can be directly used. Through a wrapper toolkit DAFFODIL can access SRU/SRW, Z39.50 and OAI data sources. Besides them access to any web based digital library is possible. DAFFODIL currently support access to the domain of computer science and can be used under <http://www.daffodil.de>.

Establishing Data Grids for SHAMAN

The goal of SHAMAN is to setup a preservation solution based on data-grid technology. Distributed data-grid technology is used to manage and administer replicated copies of digital objects over time. Data-grid middleware will be used as core data-management technology, mediating between SHAMAN components and legacy systems. Such systems are SRB and iRODS, which we will take under close evaluation, since they are widely used systems.

SRB

The Storage Resource Broker (SRB) is a grid middleware, developed at the San Diego Super Computer Center as commercial product. SRB enables integration and transparent use of different geographically distributed storage systems. A user accessing a digital object is not aware of the current location. A SRB system consists of several zones. A zone itself is represented by an arbitrary number of SRB servers and a central database, called MCAT. A SRB server can manage several storage systems (resources). Besides metadata, the MCAT also stores information about the zones, locations and resources. Clients can access via any SRB server all objects in a zone. The query is automatically routed by the MCAT. Archived objects can be structured in collections and sub collections. Another important aspect is the fact that collections can contain objects from geographically distributed sides in one logical view. Around SRB exist several so called drivers which enable access to other storage systems, e.g. GridFTP in both directions, to access SRB storage from GridFTP and vice versa. Also DSpace can integrate SRB as a storage space. SRB based collections currently hold more than 150 million files worth > 1000 TB of data.

iRODS

iRODS, the integrated rule-oriented data system, is the open-source successor of SRB, also developed by the San Diego Super Computer Center. iRODS contains the same functionality as SRB but, as new feature, introduces a rule engine. Such rules follow the event-condition-action paradigm and run on the iRODS servers together with so called micro services. Micro services can be implemented and integrated via a plug-in feature in iRODS, so there are no limitations on functionality and extensibility. Examples for such micro services are to create a copy of an ingested object or check an object for integrity based on checksums.

Furthermore micro services can then be connected to more complex rules, which can follow again events and conditions.

iRODS is already on the way to being used as a preservation system. In some institutions it is also currently in the migration process, where they change the system from SRB to iRODS.

Requirements for SHAMAN's ISP-1 Integration Concept

In order to design a first integration approach of the identified legacy systems in SHAMAN we will analyze an initial example application-scenario provided by DNB. The scenario is as follows: A memory institution uses a specific long-term preservation system for physical printed books and journals. This system is not intended to be replaced by a new system. But the system is not well equipped for handling digital objects, like web pages, which by law have to be preserved, too. The idea is now to extend the preservation solution through new technologies like a grid-based system, in order to cope with the amount of stored digital information objects. The legacy system will remain to be in use as main preservation system, but access, ingest, and management should be possible in parallel through the grid system with one interface and a set of appropriate internal workflow processes.

Within this scenario, we were able to identify a first set of three initial use cases to integrate the existing system with our grid-based system. These use cases are central access on distributed repositories, central storage on distributed archiving and central management on distributed collections.

In order to analyze, discuss, model, and later implement the three use cases, we need to specify an integration architecture. Therefore, as a first exercise towards building the overall framework's reference architecture, a specific integration architecture for ISP-1 has to be derived. This will be a starting point for the extension and abstraction of this architecture into a more general framework architecture that can serve all three ISPs and in the ideal case support many other future system developments for other application domains and scenarios as a development and deployment framework for DP application solutions.

Such an initial architecture for ISP-1 needs to fulfill certain requirements. One set of requirement is provided explicitly in the SHAMAN project plan. The SHAMAN project requires to establish a very dynamic framework for the development of a stable and reliable preservation environment which is strongly driven by supporting infrastructure independence, the ability to preserve digital entities as a collection, and the ability to migrate the collection to new choices of storage and database technologies.

In addition, the current status of requirements that are driven by all the described legacy systems is as follows:

- The data will be stored in distributed repositories. If customers integrate their systems with new technology data will continue to be stored in distributed repositories.
- The repositories are running on different legacy systems. Therefore, future distributed repositories to be established should still be able to run on different legacy systems, like DSpace, Fedora, KOPAL, or traditional database systems, too.
- The legacy systems provide different protocols. Each future system should be able to provide different searching and browsing protocols, too, as well as different protocols and processes for ingestion.
- The legacy systems use different metadata standards. Therefore, a future system should be able to support these metadata standards, such as Dublin Core, METS, MARC or LMER, too.

Utilizing Service Orientation in the Integration Architecture

The above described requirements make it necessary to use a service-oriented architecture (SOA) because it provides the following features:

Modularity - The upcoming system need to be modular to integrate each legacy system.

Standard - The system needs to standardize the protocols and metadata formats.

Independence of technology - The preservation process should not rely on any technology; it should rather be able to easily adopt new technology for better performance.

Flexibility - Each part of the system should be easily replaceable or adaptable to new needs and future technology.

Reuse - Already existing service should be reusable in other context.

In short, we need to setup a service-oriented architecture in order to provide a modern, agile, flexible and dynamic system to optimize all processes within a long-term preservation environment. Existing services can be reused and new features can be adopted and integrated without disturbing running processes. In this way it will be possible to manage such a feature-rich, complex, and dynamically evolving set of tasks as preservation solutions are faced with. This service-oriented draft of an integration architecture for ISP-1 will be a starting point or the extension and abstraction of this architecture into a more general framework architecture for the whole SHAMAN project.

In the following sections each use case is depicted by a four layered service-oriented architecture. The lowest level *Preservation/Storage Systems* holds all legacy systems as well as the grid-based repositories. The *Wrapper* level enables standardized access to the underlying systems. The *Service* level combines functionalities which represent the workflows and processes necessary to run a preservation system. On the top level the *User and Management Interface* provide users and administrators with access to the system.

Information Integration based on a Mediator Approach

In a distributed environment we need to search and browse several distributed repositories in order to support user queries. If the environment consists of more than one legacy system, a mediator or wrapper is necessary, if it is not possible to directly integrate a legacy system into the grid system. This is e.g. the case with the displayed iRODS driver for the system DSpace.

A multi-layered architecture for such an iRODS driver case

Figure 1: System Integration with iRODS Wrapper

is depicted in *Figure 1*. The legacy systems are located on the lowest level. Via iRODS wrappers/drivers we gain full access on the bases of the iRODS protocol to serve the search and browse queries. The service can rely on defined protocols and propagate the query and gather the results to be presented via the user interface within DAFFODIL. Through these mediator levels, users gain transparent read access to any legacy system.

The (read-only) search and browse process can be described the following way:

1. The user interface of DAFFODIL relies on a specific Search and Browse service and passes any query via the communication platform of the SOA to these services.
2. The service connects to the iRODS MCAT server and if
 - a) a central search index exists, runs the query central
 - b) a distributed search index exists; the query is passed to each repository and performed locally

Figure 2: System Integration with General Wrapper

3. The resulting objects will be accessed by the iRODS driver from the legacy system, e.g. KOPAL's knowledge base and passed through the services to the user.

During this process the syntactical heterogeneity of the metadata is captured on the wrapper level, whereas the semantic heterogeneity of the different search and browse interfaces is captured on the service level.

If we do not want to rely on iRODS only as long-term preservation storage system, it is also possible to abstract from it by implementing general wrappers to access any legacy or grid-based system. The above described process still holds also for this case, but the difference is, that each wrapper has to implement the search and browse functionality formerly provided by the iRODS driver as depicted in *Figure 2*.

As stated in a previous section, a first prototype implementation of this scenario can be completely based on the existing DAFFODIL framework utilizing at the same time a service-oriented architectural approach. On the lowest level we need to implement wrappers for the DNB, SUB and the FP. If they exist, the search and browse functionality is ready to be evaluated. The search service already combines results from distributed heterogeneous data sources and the user interface presents the result directly with query term highlighting, sorting and filtering. It is out-of-the-box possible to store found result in the personal library and many other already existing high-level functions can be used. Within [Klas2007] it was also proven, that the DAFFODIL system raises efficiency and effectiveness of the user during the search and browse process over any other search system.

Distributed Ingestion

Even if the archives of the DNB, SUB or FP are integrated into the grid based system archiving of new objects still takes place in the local repositories. However, the grid system needs to be aware of changes in the local repositories in order to support search and browse functions. The situation is depicted in figure 3. To overcome this problem three solutions can be discussed:

disaster, could be implemented through a replication service. Based on risk calculations and worth of the digital objects, the user states the requirements, to have three copies of the objects in distributed sites.

Whereas in the cases above, the wrapper and services need only to be aware of their local environment, in this case a

Figure 3: Distributed Ingestion

- Local ingestion and a redundant grid ingestion: Here the local system runs their ingestion process and after success runs the ingestion on the grid system. This is the least complex integration.
- Local ingestion and notification to grid server: The second case ingests also to the local repository, but either sends out a notification message, that a new object was ingested to the grid system or the grid system polls on schedule for new objects in the local repository, e.g. OAI harvesting could be used.
- Grid ingestion and triggered local ingestion: In the third case, the more un-trusted case by the local repository owner, the object is ingested in the grid system and then locally ingested.

In any of the above cases it has to be discussed where the real object is stored. Either it is stored in the local repository and only the metadata information is published on the grid, or the object itself is replicated on the grid. On the management layer the repository manager has to be always aware that the ingestion process was correct and that the integrity and authenticity of the objects is guaranteed.

The quality of this service is different to the search and browse case, since we need write access and access to the access rights management.

Managing Distributed Collections

Besides the integration of information and the distributed ingestion process, managing the distributed collections in the heterogeneous grid environment with all legacy systems is another important challenge. The management is necessary to cope with formulation and implementing of policies, prioritizing and planning, assessing risks or calculating expenses.

The use case here, holding several copies of an object on distributed repositories in order to avoid loss through e.g. a

Figure 4: Managing Distributed Collections

complete new mediator level needs to be aware of all repositories to find another repository which meets the requirements to replicate the object at that side.

In *Figure 4* the management tool within DAFFODIL initiates the replication process, after the replication policies for the user are changed.

1. The task “replicate the object x from resource KOPAL to resource DSpace” is handed to the replication service.
2. The service checks the DSpace repository, if it is available, has enough free space, etc.
3. The service initiates the copy process, which of course contains verification processes, e.g. via MD5.
4. Both repositories have then to indicate if the copy process is completed and correct which is visualized in the management interface. This indication is also logged for legal issues.

The management tool on the interface level will become a master control station in order to monitor processes, policies and archive requirements.

Outlook

Combining the first architectural models from the above three use cases, we can derive a multi-layer conceptual system model based on a service-oriented architecture, as depicted in *Figure 5*.

On the lowest level the preservation and storage systems are located. The main system in the SHAMAN context will be a grid based system. All the existing functionality of this system will be verified and reused. In case of heterogeneity problems, either of syntactical or semantic nature will be handled on the wrapper and service layer. The wrapper layer integrates and enables access to the storage systems. The service layer then supports all necessary functionality not provided by the storage systems. On the top level the user/administration/management interface relies on the lower level to visualize the complex functionalities.

Each functionality is represented by a specific communication protocol and described as a set of services with input/output parameters within the SHAMAN service oriented architecture. In figure 5 the three protocols search and browse, ingestion and management are related to ISP-1, whereas Design and Engineering as well as eScience are related to ISP-2 and ISP-3 within SHAMAN, where the necessary protocols still have to be identified.

The SHAMAN goal to define a The Theory of Preserva-

Figure 5: Multi-Layer Model on Service-Oriented Architecture

tion can be supported on this conceptual level and we will aim to prove its assumptions based on these services and protocols. The services and protocols define the SHAMAN system and in order to run a future system, a service provider only needs to be compliant to the service description and protocols. Legacy systems need to fulfill only a minimal set of services and protocols in order to be integrated or migrated into our SHAMAN system or they need to have open interfaces to be wrapped, if a customer wants to use their proven system. In order to be compliant with the SHAMAN framework other systems need to implement the necessary SHAMAN services and protocols.

Summary and Next Steps

In this paper we have described the SHAMAN project, its aims and challenges. Within the ISP-1 we identified the need to incorporate legacy systems, since some customers will not necessarily change their local running preservation environment, but need to extend and integrate new tech-

nologies to scope with future requirements. In three realistic use cases we identified challenges that we have to meet. In order to enable these we propose a sophisticated service-oriented architecture based on a multi-layer conceptual model. Doing so, we meet the above stated requirements of modularity, standards and independence from technology. Furthermore this will enable SHAMAN's demonstrators to become independent of any future preservation system, but to fulfill the needs of its users to preserve important information. Going on from ISP-1 to the whole SHAMAN project with the other domains of Design and Engineering as well as eScience we will investigate their needs in order to integrate their requirements and extent, adopt, remodel, and verify this architecture. Furthermore, the next steps to setup and evaluate the grid based SHAMAN system will be:

1. Enabling search and browse functionality on the repositories of the DNB, the SUB and the FP based on the DAFFODIL system. We will reuse existing wrapper and service implementations from previous projects where e.g. The European Library and DNB were projects partners.
2. Integration of institutional repository software DSpace and Fedora, the preservation system KOPAL and iRODS on the wrapper level as storage systems within the DAFFODIL framework in order to implement the graphical management tools and services for the ingestion process
3. Model, implement and setup management functionalities for policy processes as addressed e.g. in third use case for replication.

The best practices gained from these implementations will be evaluated and form impact on the SHAMAN overall conceptual model.

Acknowledgments

Special thanks goes to Jose Borbinha, Jürgen Kett, Alfred Kranstedt, Adil Hasan and the SHAMAN consortium for the discussions and comments. This paper is supported by the European Union in the 7th Framework within the IP SHAMAN.

References

- Fuhr, N.; Klas, C.-P.; Schaefer, A.; Mutschke, P. 2002. *Daffodil: An integrated desktop for supporting high-level search activities in federated digital libraries*. In Research and Advanced Technology for Digital Libraries. 6th European Conference, ECDL 2002, 597–612. Springer.
- Klas, C.-P. 2007. *Strategic Support during the Information Search Process in Digital Libraries*. Ph.D. Dissertation, University of Duisburg-Essen.

Creating Trust Relationships for Distributed Digital Preservation Federations

**Tyler O. Walters (Georgia Institute of Technology) and
Robert H. McDonald (Indiana University)**

Georgia Institute of Technology, Library and Information Center
704 Cherry Street, Atlanta, GA 30332-0900
Indiana University, Herman B. Wells Library 234
1320 East 10th Street, Bloomington, IN 47405-3907
{tyler} at gatech.edu, {robert} at indiana.edu

Abstract

The authors outline a model for digital preservation federation based upon several existing models including the U.S. Federal Reserve Bank regional governance model and its similarities to successful large-scale redundant internet networks. In addition other trust models will be examined including Maister, Green, and Galford (2000), Holland and Lockett (1998), and Ring and Van de Ven (1994). These models provide key frameworks for understanding how trust can be enabled among federated but independent institutions.

Introduction

As more research, educational, and cultural institutions come to realize the enormity and complexity of work required to store, preserve, and curate large amounts of their unique digital information, many will turn to establishing cooperative partnerships for leveraging existing mass-storage capacity or utilizing 3rd party data curation service providers to help satisfy their needs for a redundant and secure digital preservation system. The concept of trust and its manifestation between institutions as an essential element in designing digital preservation systems – both technical and organizational – is critical and appears in the organizational level needs of the CRL/NARA-RLG Trustworthy Repositories Audit and Certification (TRAC): Criteria and Checklist. Trust can be defined simply as “relying upon or placing confidence in someone or something...” (www.dictionary.com). With regard to preservation in digital libraries and archives, trust means that we rely upon the organizations or institutions maintaining the digital library or archives to sustain the information deposited in it, and that this information remains authentic, reliable, and unchanged over time and across technologies. We trust that the institutional actions taken upon the digital library and the content held can be

trusted to serve these goals. To achieve this, as we look at partner institutions who are participating in preserving our own institution’s digital content, we are seeking to answer whether or not their actions with our material are trustworthy. Trust is always an underlying, critical factor impacting the success or failure of inter-institutional relationships. The concept of trust is imbued in everything we do as digital library and archives professionals, especially in an inter-institutional, cooperative setting.

Increasingly, federations of institutions and organizations are being formed to devise strategies and systems to preserve digital information. The choice of the word “federations” is significant because it aptly describes what these institutions are doing. “Federation” can be defined as “people, societies, unions, states etc. joined together for a common purpose.” “...a federated body formed by a number of nations, states, societies, unions, etc., each retaining control of its own internal affairs.” (www.dictionary.com). According to these definitions, a federation is unique in that the individual institutions comprising it continue to “retain control of its own internal affairs,” while at the same time they are coming together to solve a common need. The phrase “distributed digital preservation federations” is being used increasingly to describe cooperatives of geographically-dispersed institutions who are banding together to form solutions to the digital preservation problem. Identifying and analyzing successful federation models as well as human practices that foster inter-institutional trust development are salient to the work of building distributed digital preservation federations.

Existing demonstrations of cooperative trust as well as literature on trust relationships offer much to the international digital preservation community. One successful model – the U.S. Federal Reserve Bank (Fed) regional governance (trust federation) model – stands as an exemplar for centralized authority while providing for

distributed independent organizational governance; a key concept for any digital preservation federation. The Fed has many similarities to large-scale redundant internet networks and provides key elements for sustainability of a federated organization of independent agents. Inside a cooperative, inter-organizational model of trust, there are independent institutions and the people they employ that communicate, interact, and make decisions. The literature on organizational trust can illuminate the institutional qualities its people must foster to develop successful trust relationships. Therefore, we will explore the governance framework of the U.S. Federal Reserve Bank system as well as the trust models identified by Maister, Green, and Galford (2000), Holland and Lockett (1998), and Ring and Van de Ven (1994), and which apply to the dynamics of trust and trust-building between and among separate governing institutions, and adapt them to the distributed digital preservation federation context.

Concepts, Models, and Frameworks for Trust

Within a trust model such as that of the U.S. Federal Reserve System (central banking) model posited in this paper, people, organizations, and the inter-institutional federations between them must have a formal mandate for “trust.” This type of formalized trust has been previously identified from both a contractual (Berman et. al.), evidence based methodology (Ross and McHugh), and organizational structure analysis (McDonald and Walters, 2007). In order for this trust model to succeed when applied to coordinated or federated digital preservation organizations, each autonomous entity must receive adequate preservation services while retaining appropriate autonomy for its primary institutional organization. The authors will delve further into examining what institutional and personal characteristics, principles, and building blocks must be present to foster and sustain trust in an inter-institutional model such as digital preservation federations. They will describe and discuss the dynamics of such a model and principles for building strong organizational relationships while describing the stages and key elements involved in establishing a long-term federated trust.

U.S. Federal Reserve System. The U.S Federal Reserve System is composed of twelve Federal Reserve Districts (see Figure 1), each of which has a Reserve Bank. The Federal Reserve Banks operate under the general supervision of the Federal Reserve Board of Governors which is located in the District of Columbia. While each district generates its own income from interest earned on both government securities and priced services for financial institutions, no district can operate for a profit. All profits are returned to the U.S. Treasury thus enabling a symbiotic relationship between the individual districts and its centralized governance body, the Board of Governors (Grey, 2002).

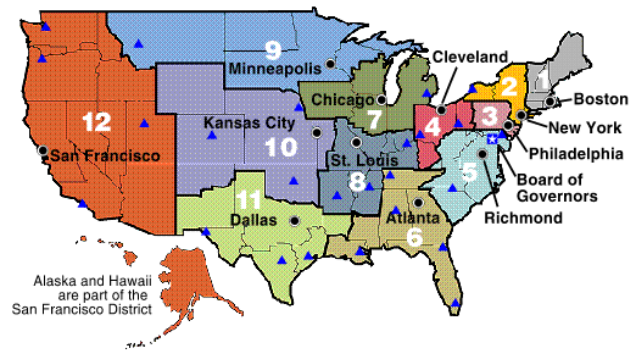


Figure 1: Map of the U.S. Federal Reserve Districts.

While this framework is somewhat artificial due to the restraints imposed upon it by the U.S. Legislative Branch, it does have one key feature that was embedded in its creation and that is the one of regional self-governance. After the failure of the 1st and 2nd Banks of the United States, the U.S. Legislative Branch wanted to build an entity that was not wholly controlled by the banking industry but that could affect central control over the economy in order to prevent disastrous short-term financial failures. The Federal Reserve Act of 1913 authored by Congressmen Glass and Owen did just that by creating a system that represented the interests, regionally of the banks that it regulates, of the United States by creating a large system of regional banks (eventually twelve-with several branches in some districts) which would have control over a central economy and yet have representation from every banking region of the United States (Cornell LII, 2008).

These districts each have their own governance based in a board of directors which is divided into three classes for representation including Class A, Class B, and Class C directors. Class A and Class B directors are elected by the regional banks of the individual Federal Reserve District while Class C directors are appointed by the System Board of Governor’s in DC. Thus a distributed system that meets the needs of local banks while implementing central stability from its System Board of Governors. When the system was created it was widely known that one of the main reasons for the failures of the 1st and 2nd Banks of the United States was that the banks were located in close proximity to the U.S. Congress and thus could easily be manipulated for political reasons. By creating a system that had both a central authority as well as regional autonomy the U.S. Congress enabled a sustainability model that is inherent in many areas of current society as derived from other large-scale autonomous systems such as that of the commercial Internet.

If we adapt this model to a distributed digital preservation bank or long-term data bank we see that for reasons of scale it will be necessary to have national and international partnerships; however, in order to retain digital works which have regional and local significance, a strong regional cooperative is needed. Both the MetaArchive Cooperative as well as other regional cooperatives such as the Alabama Digital Preservation Network (<http://www.adpn.org>) and the Committee on Institutional Cooperation's (CIC) HathiTrust (<http://uits.iu.edu/page/awac>) meet this criterion. While both the MetaArchive and the HathiTrust are actively building national and international alliances, it is the local and regional selection of content that will build strong preservation nodes over time. This in effect will give our long-term preservation partnerships regional self-governance while enabling trusted relationships for shared data curation for expertise and scale that will ensure long-term sustainability for our most precious record of knowledge.

Holland and Lockett. In the first model examined here for transactional based trust relationships we have identified one set forth by Holland and Lockett which looks at virtual organizational models. The prime motivator in this model is the idea of business and commerce being motivated by many complex partnerships in the supply chain in order to conduct business at a global scale. Much like the types of international trust relationships that digital preservation cooperatives seek, this virtual environment is built upon indicators of trust. In Figure 2 we see the trust antecedents for a collaborative federation with mapping to the organizational and individual behavioral elements and their outcomes. This mapping is typical of many non-profit virtual organizations but in this case uses communication feedback from the NARA/RLG Trusted Repository Audit Checklist and the DRAMBORA framework for trusted repositories for indicators of trust certitude.

Holland and Lockett devise five hypotheses which will be telling in the long-run as to how effective virtual organizations can be in managing national and international preservation efforts. These hypotheses are as follows (Holland and Lockett, 1998):

Hypothesis 1: Virtual organizations will develop quicker and easier where the level of subjective trust between the different economic partners is high.

Hypothesis 2: The importance of subjective trust in determining the success of virtual organizations is contingent on the risk of failure and the importance of the outcome.

Hypothesis 3. Shared information systems amongst economic partners involved in some form of virtual organization will serve to speed up the trust/ distrust development process.

Hypothesis 4. International differences in dispositional trust will become less important than situational context in determining the level of subjective trust as shared information systems enable the free flow of performance information between separately owned economic partners.

Hypothesis 5: In business markets, virtual organizations will be characterized by long-term relationships and stability rather than transient relationships to support unique projects or electronic markets.

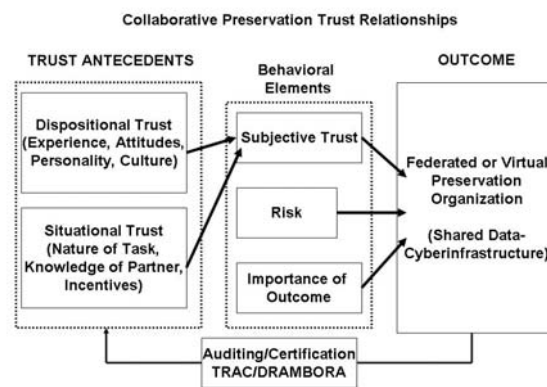


Figure 2: Adapted from Holland and Lockett Model as diagrammed for digital preservation federations.

Ring and Van de Ven. This early (1994) model is designed to examine cooperative inter-organizational relationships (IORs) and the frameworks they utilize in formal, legal, and informal social-psychological processes when negotiating and executing their business activities. Ring and Van de Ven further focus upon and explore how and why cooperative IORs emerge, evolve, and dissolve. They assert that their findings enlighten our understanding of the transactional cost economics of business being conducted through cooperative IORs as well as other aspects of business relationship development. Their modeling can help in understanding the characteristics of digital preservation federations' lifecycle stages as these efforts are initiated, ascend, and mature.

Some of the key relational phenomena Ring and Van de Ven target for study involve the balance between certain relationship parameters. Among these are: positive versus negative framing of a situation between partners; personal versus business role relationships that drive the IOR; and psychological contracts (a compatible perspective that is shared between two parties, and therefore, a positive "connection" forms between them) versus formal, documented contractual agreements. They also note that the length of time an IOR continues versus the length of

time the original persons are involved dictates when informal processes become formalized. It is this balance between formal and informal, or the lack thereof, that are indicative of successful (balanced) or failing (imbalanced) cooperative inter-organizational relationships. Where the NARA/RLG Trusted Repository Audit Checklist assists with documenting and formalizing trust relationships (NARA-RLG, 2007), the work of Ring and Van de Ven supplements our understanding by illuminating the dialectical relationship they espouse between formal and informal trust markers.

Much as Holland and Lockett did later in 1998, Ring and Van de Ven developed a seven-proposition model that identifies the characteristics of IOR initiation, growth, and dissolution:

Proposition 1: Congruent sense making among parties increases the likelihood of concluding formal negotiations to a cooperative IOR.

Proposition 2: Congruent psychological contracts among parties increases the likelihood of establishing formal commitments to a cooperative IOR.

Proposition 3: If the individuals assigned to a cooperative IOR do not change, personal relationships increasingly supplement role relationships as a cooperative IOR develops over time.

Proposition 4: Informal psychological contracts increasingly compensate or substitute for formal contractual safeguards as reliance on trust among parties increases over time.

Proposition 5: When the temporal duration of inter-organizational relationships is expected to exceed the tenure of agents, informal understandings and commitments will be formalized.

Proposition 6: As the temporal duration of a cooperative IOR increases, the likelihood decreases that parties will terminate the relationship when a breach of commitments occurs.

Proposition 7: When significant imbalances between formal and informal processes arise in repetitive sequences of negotiation, commitment, and execution stages over time, the likelihood of dissolving the cooperative IOR increases.

Dynamics of Trust

While international, national, and regional frameworks and models need to be pursued, institutions also need to grow their abilities in developing trust relationships between one another. In order to advance these relationships, we need to examine the qualities and characteristics of trust relationships within the context of organizational learning

and behavior. Authors Maister, Green, and Galford explore this in the book, *“The Trusted Advisor”* (2000), which can be adapted and applied to building successful models for trust relationships in distributed digital preservation federations. While the authors focus on professional services personnel acting in the role of advisors to companies – the advisor / client relationship – they articulate useful models of trust development, provide many insights based on their experiences as organizational consultants, and identify desirable organizational qualities for successful inter-institutional relationships, such as those we find in distributed digital preservation federations. Their work will be examined in the following sections and applied in a cursory way to the early experiences in federation-building for distributed digital preservation.

Trust and the Individual. Perhaps the most major insight offered by Maister, Green and Galford is that institutional trust isn't institutional at all. Trust is built between individuals (working for the institution); therefore people in institutions grow trust between them, and then bring their institutions into partnerships based on that trusted relationship. To go forward, there must be a satisfactory level of assurance that each institution will perform their roles and responsibilities for the other. Trust is built organically and based upon the experiences each institution has with the other. This is stated by our trust experts Maister, Green, and Galford as, “trust results from accumulated experiences, over time.” (p.23). Thus these observations, known in aggregate as institutional trust, take on human qualities because it is established and maintained by people. It is both rational and emotional, as people are. The emotional side is something we must pay close attention to if trust relationships are to flourish. For instance, we value trusted colleagues when they comprehend, support, and demonstrate a dedication to achieving objectives that are complimentary to our own institution's objectives. Colleagues may not always agree and they may even challenge our viewpoints. However, they do so with care, and maintain a concerted sense of achieving the shared objectives. Therefore, we trust their motives and lines of questioning. In this scenario, our emotional self initiates and we ask ourselves questions like:

- Does this colleague understand me, or is she pushing her own agenda?
- Is she helping me think through a problem, or is she just trying to substitute my thinking for hers?
- Does she have my interests at heart, or her own? Is she on my side?
- Am I comfortable with her style, or is she overbearing and domineering?

- Is she giving me new perspectives to consider, and is she doing it in a way that I'm comfortable with?

These are examples of the internal questions we ask ourselves as we assess and evaluate whether or not a colleague – and by extension, her institution – can be trusted. If the answer to many of these questions are “no,” then we conclude that the colleague does not share our objectives and viewpoints. We immediately question her motives and ultimate goal. Asking ourselves, do we trust her? Can we trust her words and actions? The decisions we make about individuals with whom we enter into business relationships is extremely personal and this process applies to the world of building distributed digital preservation federations as well.

Establishing Institutional Trust. With trust being a human-based rational and emotional process, as well as a process of accumulation and growth, one can conclude that trust relationships are a “two-way street.” Trust is fundamentally about assessing and managing the risk perceived by each institution entering into a relationship. In other words, “trust entails risk” and is thus one of the components of any trusted federation (p.24). Any partner in a trust relationship can choose to either follow through on the agreed upon actions, or do something different. However, because of the trust relationship it is most likely the partners won't do something different. (p.24). This is due to the nature of trust relationships where the institutions involved both participate (i.e. “get”) as well as reciprocate (i.e. “give”) in the relationship. Neither wants to upset the balance, otherwise the equation falters and the collaboration is no longer of benefit.

Maister, Green and Galford, posit that there exists a “trust equation” expressed as:

$$\text{Credibility} + \text{Reliability} + \text{Intimacy} / \text{Self-Orientation} = \text{Trustworthiness}$$

These four primary components bear examination as we attempt to establish successful and long-lasting distributed digital preservation federations.

Credibility. Both credibility and reliability are the most tangible of the four components. Credibility comes from the mastery of our professional body of knowledge and how we communicate it. Therefore, credibility has both rational and emotional elements. Maister, Green, and Galford state that credibility is content expertise plus “presence,” referring to how we look, act, react, and talk about our content.” It depends not only on the substantive reality of the advisor's expertise, but also on the *experience* of the person doing the perceiving.” (p.71). This relationship illustrates the “two-way street” paradigm of

trust relationships. To build credibility, it is not only about expertise; it is about how that expertise is communicated and then perceived by the person receiving it. Credibility is about words and language, including non-verbal language.

Reliability. If credibility is about the use of language to communicate expertise, then reliability is about the actions taken to fulfill a promise or intention that was communicated. It is about “the repeated experience of links between promises and action” (p.74) or “of expectations fulfilled.” (p.75). Creating opportunities to demonstrate reliability to prospective partners is best done “by making promises, explicit or implicit, and then delivering on them” (p.75). Here too with reliability, there are rational and emotional aspects. The emotional aspects relate to doing things in ways that our partners are familiar with and prefer. Therefore, our own institution's culture needs to support learning about our partners with whom we conduct business, their preferred ways of “doing business,” and then deliver on our promised roles and responsibilities in ways they are accustomed. As new federations of digital preservation activity arise, we must recognize that reliable, dependable behavior by our institutions may not be perceived as such by our partners. We must understand this and learn how they perceive and measure reliability in a partner's actions, then set out to behave in recognizably dependable ways.

Intimacy. Intimacy and self-orientation are the more elusive of the four trust components. Intimacy refers to our emotional response to words and actions. It is about our intuitions in regards to who we are interacting with and whether or not we are comfortable in this interaction. Describing intimacy, Maister, Green, and Galford offer:

“People trust those with whom they are willing to talk about difficult agendas (intimacy), and those who demonstrate that they care (low self-orientation).” “Intimacy is about ‘emotional closeness’ concerning the issues at hand... it is driven by emotional honesty, a willingness to expand the bounds of acceptable topics, while maintaining mutual respect and by respecting boundaries. Greater intimacy means that fewer subjects are barred from discussion.” (p.77).

In digital preservation partnerships we need to achieve a state where collaborators from different institutions can challenge each other's thinking, take each other to task on comments made, be critical (constructively, of course), and be very honest about difficult matters as they occur. Strong emotions may arise and they need to be communicated, while the others receiving this emotional communication need to be comfortable enough to allow these expressive moments to continue and resolve themselves. In these cases, people only need to convey their thoughts and be validated that they have a certain point of view, as opposed

to changing everything because of that view. Once these experiences occur and everyone accepts what was said (not necessarily agreed to), intimacy develops and people feel that more topics can be discussed and resolved. There will be many lurking, hidden issues to resolve in delicately balanced, broadly-based preservation federations. Increasingly, these federations could be international in their composition. The more intimacy developed between partners means the more they will examine tough issues, discuss, and resolve them, all to the benefit of the federation's operations.

Self-orientation. Self-orientation is a critical concept that can make or break the success of any federation. It is about this sense of giving to others that permeates all collaborative work. If an institutional partner feels that another partner is being self-serving and not considering the needs of the other partners, and then their motives are questioned, they are not trusted, and eventually they are marginalized or perhaps even removed from the federation. Maister, Green, and Galford, on self-orientation, state that "there is no greater source of distrust than advisors (i.e. partners) who appear to be more interested in themselves than in trying to be of service to the client (i.e. the other partners)" (p.80). Further, "...any form of preoccupation with our own agenda is focusing on something other than the client (i.e. partners), and it will reduce trust directly" (p.81). Several steps can be taken to build core values into our organizational cultures that value understanding our partner institutions. Some of the inter-personal abilities to be cultivated in preparing a "partner-ready" organizational culture are (pp. 80-81):

- Recognizing that "defining the problem" is the most important activity, as opposed to being the institution that initiates the plan or technique to solve the problem.
- Listening actively to one another, summarize what is being heard from your partners.
- Discussing the motivators behind an issue, not just discussing the issue itself (this requires intimacy).
- Being willing to say "I don't know" when we truly don't know (shows authenticity, builds credibility).
- Acknowledging each other's thoughts and feelings on a given topic.

Focusing on what others are expressing is critical to lowering self-orientation, which supports staying focused on partner needs and, in turn, builds trust in the relationship. If the institutions in your federation truly share the same problem space, and you've done the work of selecting partners correctly, then what is good for them will be good for your institution as well.

Maister, Green, and Galford assert that the "trust equation" is not "just so much softness" (p.83), but rather it has real consequences for the economic costs of business relationships. Costs go down if business can be generated with existing clients because trust relationships have been formed. The authors conclude "the cost of developing new-client business is 4 to 7 times higher than the cost of developing the same amount of business from an existing client" (p.84). Similarly, with digital preservation federations the costs of developing new trust relationships is high. Federations like the San Diego Supercomputer Center's Chronopolis Project and the MetaArchive Cooperative's MetaArchive of Southern Digital Culture both began by working with partners from previously existing multi-institutional projects, each which had an interest in digital preservation. The major motivations to federate were: 1) the desire to hold down costs (a shared value and scale); and 2) to find partners around which they could build a trust relationship to advance a new, complex preservation federation. This meant finding institutions with whom they had already *invested* in a trust relationship. This was one way of reducing costs, advancing the federation quickly, and with high-quality outcomes. Developing trust relationships costs time, effort, and resources. Models for building them such as the trust equation helps us identify proper modes of conducting our "preservation business" inter-personally and inter-institutionally.

Advancing Trust Relationships

Balancing the components of the trust equation and the inter-personal abilities that have us focusing on our partners' needs while meeting our own institution's objectives, may seem counter-intuitive. It feels like an act of faith, trusting that our partners will put our own institution's objectives in the forefront. To further illustrate how an ascending cycle of trust grows to enable the trust relationship phenomenon, Maister, Green and Galford, identify and describe five stages in the development of trust. They are: 1) Engage; 2) Listen; 3) Frame; 4) Envision; and 5) Commit. Their work focuses not on "solving the problem," but rather on "building the relationships" that keeps institutions together who will eventually solve the problem.

A cursory understanding of these stages will help us to see their impact on building distributed digital preservation federations. The "Engage" stage establishes that partners have identified an issue worth discussing, and that they are worthy institutions to discuss the matter with, given their adequate desires or expertise regarding digital preservation. Second, is the "Listen" stage, where partners believe they understand one another's perspectives, experiences, and approaches to digital preservation. This third stage known as "Frame" is when one or more partners

help “crystallize and clarify the many issues involved” (p.87) in the digital preservation problem for another partner. The receiving partner realizes that value is being added by the clarifying partner; hence a significant amount of trust can be developed in this stage. The fourth stage, “Envision,” we are not yet offering solutions to the problem of digital preservation. Instead, this stage is when partners join together and develop options for how the problem may be resolved. This is a visioning period where many approaches are imagined. Together the partners begin to better understand their goals and what is required to meet them. The fifth and last stage is “Commit.” This stage is where the partners understand “in all its rational, emotional, and political complexity, what it will take to achieve the vision, and to find the determination to do what is necessary.” (p.89). Commitment begets action, which is taken by the partners together as a federation to resolve digital preservation issues. Being aware of these trust development stages helps to nurture business relationships that can withstand misunderstandings and differences of opinion to band together resources, imagine new digital preservation approaches, and enact them.

Conclusions

The successful preservation of valuable digital assets will require the expertise and collaboration of many individuals and institutions, both in the public-sector as well as in the commercial sector. In order for the library, archives, museum, and the broader cultural memory sector to accomplish their goals of long-term preservation for the world’s knowledge, records, and, artifacts, it will be necessary to build collaborative partnerships both from the stand point of a regional perspective, as well as from a national, and international perspective.

This paper presents ideas for governance frameworks as well as solid business principles for developing trusted relationships both from the stand point of public and commercial entities. The scale and complexity of the issues that need to be addressed in the preservation community will require this type of self-interested governance and collaboration model in order to succeed. More work is needed to address the question of how we will build these new collaborative organizations. With successful data preservation and access as the ultimate objectives, the implementation of structural mechanisms such as formalized trust agreements as well as business modeling in relation to organizational trust development will provide the means by which we can achieve our long-term goals of preservation, access, and discovery.

Authors

Tyler O. Walters is the Associate Director for Technology and Resource Services at the Georgia Institute of

Technology Library and Information Center in Atlanta, GA. Robert H. McDonald is the Associate Dean for Library Technologies at the Indiana University Libraries in Bloomington, IN. The authors have had close associations with sustainability planning for digital preservation federations such as the U.S. Library of Congress’ NDIIPP partnership known as the MetaArchive Cooperative (<http://www.metaarchive.org>), the San Diego Supercomputer Center’s Chronopolis DataGrid Preservation Cooperative (<http://chronopolis.sdsc.edu>) and the Committee on Institutional Cooperation’s (CIC) Shared Digital Repository (<http://uits.iu.edu/page/awac>). They are actively engaged in building leading digital repositories at the national, regional, and institutional levels.

References

- Berman, F., A. Kozbial, R.H. McDonald, B.E.C. Schottlaender. 2008. The Need to Formalize Trust Relationships in Digital Repositories. *Educause Review* 43(3).
<<http://connect.educause.edu/library/erm0835>>.
- Cornell University Legal Information Institute. 2008. *12 USC 221-522*. Accessed on 15th August, 2008.
<http://www4.law.cornell.edu/uscode/uscode12/usc_sup_01_12_10_3.html>
- Dictionary.com. 2008. Definition of *trust*. Accessed on 14th August, 2008.
<<http://dictionary.reference.com/browse/trust>>.
- Dictionary.com. 2008. Definition of *federation*. Accessed on 14th August, 2008.
<<http://dictionary.reference.com/browse/trust>>.
- Grey, George B. 2002. *Federal Reserve System: Background, Analysis, and Bibliography*. New York: Nova Science Publishers.
- Holland, C.P. and A.G. Lockett. 1998. Business Trust and the Formation of Virtual Organizations. *Proceedings of the Thirty-First Hawaii International Conference on System Sciences*, v. 6: 602-10.
- Maister, D.H., C.H. Green, R.M. Galford. 2000. *The Trusted Advisor*. New York: Simon and Schuster.
- McDonald, R.H. and T. O. Walters. 2007. Sustainability Models for Digital Preservation Federations. *Proceedings of DigCCurr 2007: An International Symposium in Digital Curation*. <<http://hdl.handle.net/1853/14442>>.
- Ring, P.S. and A. Van de Ven. 1994. Development Processes of Cooperative Interorganizational

Relationships. *Academy of Management Review*, Vol. 19(1): 90-118.

RLG-NARA Digital Repository Certification Task Force. 2007. *Trustworth Repositories Audit and Certification: Criteria and Checklist*.
<<http://www.crl.edu/content.asp?l1=13&l2=58&l3=162&l4=91>>.

RLG/OCLC Working Group on Digital Archive Attributes, 2002. *Trusted Digital Repositories: Attributes and Responsibilities*,
<<http://www.rlg.org/longterm/repositories.pdf>>.

Ross, S., and A. McHugh. 2006. The Role of Evidence in Establishing Trust in Repositories. *D-Lib Magazine* 12(7/8).
<<http://www.dlib.org/dlib/july06/ross/07ross.html>>.

United States Federal Reserve. 2008. *Map of the U.S. Federal Reserve Districts*. Accessed on 15th August, 2008.
<<http://www.federalreserve.gov/Pubs/frseries/frseri3.htm>>.

United States Federal Reserve. 2008. *Federal Reserve Act Amended*. Accessed on 15th August 2008.
< <http://www.federalreserve.gov/aboutthefed/fract.htm>>.

The Use of Quality Management Standards in Trustworthy Digital Archives

Susanne Dobratz*, Peter Rödиг**, Uwe M. Borghoff**, Björn Rätzke****, Astrid Schoger***

*Humboldt-Universität zu Berlin, University Library
Unter den Linden 6
D-10099 Berlin
dobratz@cms.hu-berlin.de

**Universität der Bundeswehr München / University of
the Federal Armed Forces Munich,
Werner-Heisenberg-Weg 39,
D-85579 Neubiberg
peter.roedig|
uwe.borghoff@unibw.de

***Bayerische Staatsbibliothek / Bavarian State Library
Ludwigstraße 16,
D-80539 München,
astrid.schoger@bsb-muenchen.de

****Rätzke IT-Service
Strasse 58 Nr. 10,
D-13125 Berlin,
bjoern@raetzke.org

Abstract

Quality management is one of the essential parts to become a trustworthy digital archive. *The German network of expertise in Digital long-term preservation (nestor)*, in cooperation with the *German Institute for Standards (DIN)* has undertaken a small study in order to systematically analyse the relevance and usage of quality management standards for long-term preservation and to filter out the specific standardisation need for digital archives. This paper summarises the first results of the study. It gives a first overview on the differences in understanding the task “quality management” amongst different organisations and how they carry out appropriate measures like documentation, transparency, adequacy, and measureability in order to demonstrate the trustworthiness of their digital archive.

1 Introduction

Already in 1996, the Task Force on Archiving of Digital Information by *The Commission on Preservation and Access* and the *Research Libraries Group* called for a certification programme for long-term preservation repositories: ‘...repositories claiming to serve an archival function must be able to prove that they are who they say they are by meeting or exceeding the standards and criteria of an independently-administered program for archival certification ...’ [11]. Some investigations in creating criteria and measuring the risk for a long-term preservation of digital objects have been carried out by several stakeholders, like the ‘*Cornell Library Virtual Remote Control Tool*’ project of Cornell University[5], the ERPANET project[4], and most recently by the Digital Repository Certification Task Force of the Research Libraries Group (RLG) and OCLC, the Digital Curation Centre (DCC) in cooperation with the European Commission funded project Digital Preservation Europe (DPE) and the German *nestor* project.

The existence of such criteria led to increased conception and installation of digital archives during the last couple of years. It also created new discussions on the importance and applicability of existing standards as many of the organisational criteria in those catalogues refer to specific ISO quality management standards like ISO 9000 etc.

During the establishing of a DIN/ISO Working Group in Germany for defining criteria for trustworthy digital archives, the ostensible question on the recent degree of acceptance and usage of quality management standards within the cultural heritage sector (libraries, archives, museums) arose. Therefore the *German Institute for Standards (DIN)* sponsored a small study in order to systematically analyse the relevance and usage of quality management standards for long-term preservation and to filter out the specific standardisation need for digital archives. This study has two parts: (1) a survey amongst different digital archives and (2) an analysis of standards for the management of quality, processes, and security. It discusses the relevance and applicability in practice of those standards for use within a digital preservation environment. It shows, how and which standards related to quality management are in use in digital archives of different kind in Germany: libraries, archives, data centres, publishers, museums.

1.1 Long-term preservation and trustworthy digital archives

One of the central challenges to long-term preservation in a digital repository is the ability to guarantee the authenticity and interpretability (understandability) of digital objects for users across time. This is endangered by the aging of storage media, the obsolescence of the underlying system, the application software as well as

changes in the technical and organisational infrastructure. Malicious or erroneous human actions also put digital objects at risk. Trustworthy long-term preservation in digital repositories requires technical, as well as organisational provisions. A trustworthy digital repository for long-term preservation has to operate according to the repository's aims and specifications. Key concepts that demonstrate trustworthiness are e.g. transparency and documentation. In order to evaluate trustworthiness the measures taken in order to minimize the risk potential for the digital objects representing the important values in digital archives, have to be appropriate, measurable, and traceable.

Trustworthiness

Trustworthiness of a system means that it operates according to its objectives and specifications (it does exactly what it claims to do). From an information technology (IT) security perspective, integrity, authenticity, confidentiality, non-repudiation, and availability are important building blocks of trustworthy digital archives. Integrity refers to the completeness and exclusion of unintended modifications to archive objects. Unintended modifications could arise, due to malicious or erroneous human behavior, or from technical imperfection, damage, or loss of technical infrastructure. Authenticity here means that the object actually contains what it claims to contain. This is provided by documentation of the provenance and of all changes to the object. Availability is a guarantee (1) of access to the archive by potential users and (2) that the objects within the archive are interpretable. Availability of objects is a central objective, which must be fulfilled in relation to the designated community and its requirements. Confidentiality means that information objects can only be accessed by permitted users. Potential interest groups for trustworthiness are:

- archive users who want to access trustworthy information – today and in the future,
- data producers and content providers for whom trustworthiness provides a means of quality assurance when choosing potential service providers,
- resource allocators, funding agencies and other institutions that need to make funding and granting decisions, and
- long-term digital archives that want to gain trustworthiness and demonstrate this to the public either to fulfill legal requirements or to survive in the market.

There is a wide range of preservation archives that exist or are under development: from national and state libraries and archives with deposit laws; to media centres having to preserve e-learning applications; to archives for smaller institutions; to world data centres in charge of 'raw' data. Trustworthiness can be assessed and demonstrated on the basis of a criteria catalogue.

Documentation

The goals, concepts, specifications, and implementation of a long-term digital archive should be documented adequately. The documentation demonstrates the development status internally and externally. Early evaluation

based on documentation may also prevent mistakes and inappropriate implementations. Adequate documentation can help to prove the completeness of the design and architecture of the long-term digital archive at all steps. In addition, quality and security standards require adequate documentation.

Transparency

Transparency is achieved by publishing appropriate parts of the documentation, which allows users and partners to gauge the degree of trustworthiness for themselves. Producers and suppliers are given the opportunity to assess to whom they wish to entrust their digital objects. Internal transparency ensures that any measures can be traced, and it provides documentation of digital archive quality to operators, backers, management, and employees. Parts of the documentation which are not suitable for the general public (e.g. company secrets, security-related information) can be restricted to a specified circle (e.g. certification agency). Transparency establishes trust, because it allows interested parties a direct assessment of the quality of the long-term digital archive.

Adequacy

According to the principle of adequacy, absolute standards cannot be given. Instead, evaluation is based on the objectives and tasks of the long-term digital archive in question. The criteria have to be seen within the context of the special archiving tasks of the long-term digital archive. Some criteria may therefore prove irrelevant in certain cases. Depending on the objectives and tasks of the long-term digital archive, the required degree of fulfilment for a particular criterion may also differ.

Measurability

In some cases - especially regarding long-term aspects - there are no objectively assessable (measurable) features. In such cases we must rely on indicators showing the degree of trustworthiness. As the fulfilment of a certain criteria depends always on the designated community, it is not possible to create "hard" criteria for some of them, e.g. how can be measured, what adequate metadata is? Transparency also makes the indicators accessible for evaluation.

Recent research on trustworthy digital repositories

The ideas discussed in this paper are based on early developments on a framework describing requirements and functionalities for archiving systems that focus on the long-term preservation of digital materials, the Open Archival Information System (OAIS) [2]. From that work the Digital Repository Certification Task Force of the Research Libraries Group (RLG) and OCLC derived attributes and responsibilities for so called trusted digital repositories in 2002 [10] and finally released in February 2007, under the title: *Trustworthy Repositories Audit and Certification Checklist (TRAC)* [7], a checklist useable to conduct audits, worked out by the *Auditing and Certification of Digital Archives project* run by the *Center for Research Libraries (CRL)*. The German *nestor* project developed a catalogue of criteria in 2004 and a second version in 2008. *nestor* is concentrating on the specific national situation and elaborates the catalogue as guideline for the conception and design of a trustworthy digital archive [6]. The Digital Curation Centre (DCC) in coop-

eration with the European Commission funded project Digital Preservation Europe (DPE) conducted some test audits based on the first draft of the RLG-NARA/CRL checklist and developed a risk management tool for trusted digital long-term repositories, called *Digital Repository Audit Method Based on Risk Assessment (DRAMBORA) in 2007*[3]. Within the PLANETS project¹, the development of a Preservation Test Bed to provide a consistent and coherent evidence-base for the objective evaluation of different preservation protocols, tools and services and for the validation of the effectiveness of preservation plans takes place. In January 2007 the OCLC/RLG-NARA Task Force, CRL, DCC, DPE and nestor agreed upon a set of so called common principles, ten basic characteristics of digital preservation archives [8].

The current TRAC checklist is the basis for an ISO standardisation effort led by David Giaretta (DCC) and carried out under the umbrella of the OAIS standards family of the Consultative Committee for Space Data Systems (CCSDS) via ISO TC20/SC13.

The questions that all those standardisation efforts have to answer are:

1. Is a new single standard for trustworthy digital archives needed?
2. How does this standard refer to existing standards?
3. Is an evaluation or even a certification of trustworthy digital archives desirable and useful?

1.2 Quality management (QM) and standards

Quality of products, processes, and systems is a key factor for economical success in an open world. Implementing and operating a quality management system is vital for many organisations in order to survive on the market. But also public administrations are interested in a more efficient and effective use of revenues resources for public services. Therefore numerous principles, methods, practices, and techniques have been developed in the last decades. Many of them are consolidated, broadly accepted and published in standards.

In order to get a first idea of core concepts we refer to the well known standard ISO 9000. Quality management is defined as coordinated activities to direct and control an organisation with regard to quality. The activities generally include the establishment of a quality policy and quality objectives, quality planning, quality control, quality assurance, and quality improvement. These specific activities are the task of a quality management system. Of course, ISO 9000 also provides a definition of the term quality. It is defined as the degree to which a set of inherent characteristics fulfils requirements. And a requirement is a need or expectation that is stated, generally implied, or obligatory.

2 Background and focus of this study

The German Ministry of Economics and Technology (BMWi) has financed a long-term project called *Innovation with Norms and Standards (INS)* since 2006. The primary aim is to provide optimal business conditions for future innovation and to support their ability to act on the global market. In 2008, within the INS initiative, DIN and nestor carry out a project targeting at the standardisation of topics relevant to long-term preservation especially (1) quality management for trustworthy digital archives, as documented in this study, and (2) the standardisation of ingest processes. This project continues the work done in 2007 where the needs for standardisation in digitisation and long-term preservation have been collected and within separate studies, (1) measures within a standardised administration as well as (2) the usage of persistent identifiers have been investigated.

The present study analyses several quality management standards regarding their applicability for the evaluation of trustworthiness of digital archives. It extracts to which extent the standardisation of criteria for trustworthy digital archives can be based on existing standards and identifies domain specific standardisation needs.

Identifying and practising quality measures within a long-term preservation context attracts nationally and internationally high attention.

While the amount of digital data explodes and an growing amount of institutions are establishing digital archives, there is still a deficit in standards and commonly accepted measures used for the development and the quality control during operation of such archives.

Internationally there are two ways: first to define catalogues of criteria and second to work out risks potentials based on the specific goals of the considered archives. Thereby the links to existing standards and norms are used without defining and specifying the relation to or the use of those standards within a long-term preservation archive.

Furthermore it is useful to distinguish between the efforts towards standardisation and the efforts towards certification. The latter issue can only be carried out, if reliable standards, criteria, and most important, appropriate metrics exist.

Due to the varying goals and realisations of digital archives it is necessary to identify categories of digital archives that may use the same or similar standards.

The main focus of this study is to assess the applicability of standards. Certification methods and schemas will be subject of a follow-up study in 2009.

¹¹ <http://www.planets-project.eu>

3 Methodology

3.1 Identification of relevant quality management standards

In a first step we identified and characterised QM standards that are potentially useful for planning and operating trustworthy digital archives. Attributes already defined for determining the trustworthiness of digital archives serve as a guideline for selecting a first set of relevant standards. This first selection provides a reference in the questionnaire in order to find out easier which standards are concretely known, discussed, or already applied or refused. Moreover, this set of standards serves as a basis for a deeper analysis of the applicability of QM standards in long-term preservation considering the results of the questionnaire.

3.2 Survey of quality management standards used in long-term preservation

Second, the questionnaire and survey were designed. We asked all institutions involved in the 2004 survey on attributes and technologies used for setting up digital archives. This survey conducted by the nestor *Working Group on Trusted Repository Certification (nestor WG TDR)* finally resulted in the design of the first nestor catalogue released in June 2006.

In addition, institutions that were known to work on establishing a digital archive as well as commercial partners (e-newspapers, repository services providers) were included in the study. 53 institutions representing the digital archive landscape in Germany were asked: libraries, libraries at universities, museums, archives (public bodies), archives (private, corporate bodies), and commercial vendors.

The design of the questionnaire should mirror some of the criteria in the nestor catalogue as well as make visible those activities that could be interpreted as quality management although they might not be recognised as such by the institution. We asked for the institution's characteristics as well as for the policy of the digital long-term preservation archive and the kind and amount of digital objects hold. Several specific questions focused on the use of standards and quality management.

The 44 questions were the following²:

A	Organisation
1-6	Contact data of responsible manager <i>Information about the organisation itself</i>
7	Status of the organisation (public, private)
8	Type of organisation (administration, university, library, archive, museum, ...)
9	Research area (astronomy, biology, chemistry, ...) ³

² Details and the whole questionnaire will be given in the final study report scheduled for November 2009.

10	Mission of the institution
11	Age, growth, budget of institution <i>Information about the digital archive</i>
12	Policies
13	Growth of digital objects
14	Financial concept
15	How can the existence of the digital archive granted after structural changes in organisation? <i>Quality and security management</i>
16	Quality management (yes, no)
17	Quality management: what is done precisely?
18	Do you have a quality manager?
19	Have you concerned about standards and norms?
20	Have you discussed standards and norms?
21	Has the applicability of standards been analysed in your institution?
22	Would you need support and training in order to introduce standards?
23	Do you follow standards with a quality or security issue? (followed by a detailed list of selected standards from the theoretical analyses and by checkboxes indicating the degree of use and certification)
24	Do you follow other standards?
25	Are you developing software?
26	Do you use a service provider for the operation of the digital archive? (relation to provider)
27	Does your service provider perform a quality management?
B	Object Management
	<i>Ingest</i>
28	Types of objects (carrier, format, content)
29	Selection criteria (yes, no, planned, published)
30	Do you have formal regulations with producers?
31	Do you have a concept for keeping the quality in the relation with the producers?
32	Do you carry out quality control measures for objects and metadata? <i>Access</i>
33	Do you know your user community?
34	Have you collected the user community needs?
35	Do you provide specific interfaces for your users?
36	Do you monitor user satisfaction?
37	Do you have a concept for keeping the quality in relation to your users?
C	Infrastructure and Security
38	Have you defined the processes and organisational structures for the operation of your archive?
39	Have you documented the processes and organisational structures for the operation of your archive?

³ It was a disadvantage that no formal subject schema was used here, we oriented on a subject schema of CRL colleges.

40	Do you have an IT-concept for your institution?
41	Do you have a security concept for your institution?
42	Have you documented or contracted the commitment to upgrade your hard- and software?
	<i>Trustworthy digital archive</i>
43	Would the development of a special standard for trustworthy digital archives be helpful for your development of a long-term preservation archive?
44	Would you be interested in a certification as trustworthy digital archive? (yes, no, under which conditions?)

3.3 Applicability and practise of quality management standards

Having the results of the questionnaire at hand we can continue to analyse our pre-selected standards. Missions, tasks, and organisational forms of memory organisations as well as legal and financial constraints will allow us to determine the degree of applicability of QM standards more reliably. Therefore we have to develop a set of criteria in order to make the assessment of applicability transparent. For example, the size of an organisation or the extent of in-house software development determines the adequacy of quality standards. Of course, we additionally consider all requirements and constraints concerning QM standards explicitly stated by memory organisations within the questionnaire and related discussions.

4 Realisation

4.1 Identifying relevant quality management standards

This section illustrates how we have determined a first set of QM standards potentially useful for trustworthy digital archives.

Obviously there are several similarities between issues addressed by quality management systems and the attributes required for trustworthy digital archives.

Assessing the trustworthiness of archives needs a holistic view on the system responsible for the preservation of information. QM Systems also underpin that all components of an organisation have to be considered in order to improve quality of products, processes, and systems. Moreover, both approaches emphasise the task to investigate and respect customer needs. Therefore, we have taken generic and high-level QM standards into account. Since the preservation of digital information is highly dependent on reliable IT-systems we have also considered IT-specific standards dealing with the quality of IT on an organisational and management level.

Moreover, security is another indispensable attribute for the trustworthiness of archives. Therefore our study also comprises standards that are mainly focussing on the management of IT-systems security.

Additionally, there are many specific quality standards available. They generally concentrate on distinct characteristics of products or processes like the operating and

stocking conditions for storage media or devices. This category of standards is out of scope here, since they do not address quality management systems directly. But of course it is one of the tasks of a QM system to implement and control processes that identify, assess, and apply such standards.

These considerations lead to a first set of QM related standards that will be investigated in more detail in order to check for applicability in practise.

4.2 Survey

The survey took place during June and July 2008 when the questionnaire was distributed as PDF form and collected via email. The survey was restricted to Germany, because the financial and time resources were very limited and the purpose has been to initiate national activities.

The participants had approximately three to four weeks time to deliver the answers electronically or via fax.

4.3 Comparison of theoretical and practical results

As third step we will compare the more theoretical considerations with the answers from the survey. Since this step is still work in progress, we can only state the basic findings in this paper so far. The final report of the study is scheduled for the end of November 2008.

The goal is to investigate the usability of standards in practise and to figure out the hurdles that prevent institutions to effectively use standards. We want to find out the contexts of the standards and their portability into the area of long-term preservation.

5 First results of the study

5.1 Identified quality management standards⁴

Here we present some members of our set of identified standards and illustrate their potential usefulness for trustworthy archives.

Let us start with a glance at the popular ISO 9000 family. ISO 9000 describes fundamentals and introduces principles of quality management, which correspond to the principles and derived criteria as formulated in the nestor catalogue for trustworthy digital archives in varying degrees. Documentation, internal and external transparency and adequacy are basic principles in this catalogue. For example, ISO's quality management principles stress the customer focus, the process approach, and leadership. Leadership means to establish unity of purpose and direction of the organisation, which leads to an adequate organisational form. The process approach facilitates an integrated view to the long-term preservation of information. The customer focus corresponds primarily to the definition of the archive's designated community. The ISO standard also underpins the value of documentation. Documentation enables communication of intent, both

⁴ A first report for this phase of the study is scheduled for the end of August.

internally and externally, and consistency of action, and it serves as a mean of traceability. ISO 9000 also provides a consistent set of definitions for terms relating to quality management and introduces different types of documents used in the context of quality management. Based on the fundamentals of ISO 9000 another member of the family, namely ISO 9001, defines requirements for a quality management system where an organisation needs to demonstrate its ability to provide products that fulfil customer and applicable regulatory requirements and aims to enhance customer satisfaction. Audits are used to determine the extent to which these requirements are fulfilled. Audits can be conducted internally or externally (formal and informal). Guidance for auditing can be found in ISO 19011. With the help of a certificate an organisation can contribute to external transparency and increase confidence in its capabilities.

Maturity models are another category of standards that are useful for quality management. They define a set of attributes that facilitate to find out the maturity of an organisation to fulfil certain tasks. CMMI (Capability Maturity Model Integration)⁵ is a popular example, which has its origin in the evaluation of software subcontractors. CMMI now offers an extensive framework for process improvement and for benchmarking organisations mainly with the focus on development projects. Despite this project oriented view, we have recognised useful concepts and elements. CMMI also considers cross-project organisational aspects and, like ISO 9000, complies with the process oriented approach. Especially, CMMI stresses the institutionalisation of processes and provides generic goals and practices for the management of processes, which includes for example defining, planning, implementing, monitoring, and controlling of processes; planning of processes also covers the provision of adequate resources like funding, skilled people, or appropriate tools. CMMI additionally addresses a range of specific issues like requirements development, requirements management, or risk management as well as process and product quality assurance. CMMI also describes procedures for internal and external assessments.

Information security, primarily in the area of digital information, is another prerequisite for trustworthiness. Information security needs to be managed like quality and processes. Information is the core product of an archive. Fortunately, we can refer to already existing standards especially to the ISO 27000 series. ISO 27000 (still under development) specifies the fundamental principles, concepts, and vocabulary for the ISO 27000 series. ISO 27001 defines the requirements for an Information Security Management System (ISMS). ISO 27002 provides code of practices, for example in the areas of security policies, organisation of information security, access control, information security incident management, and business continuity management. Procedures for certification and self assessment are also addressed by this series of standards.

Of course, we have to bear in mind that these potentially useful standards are not primarily designed for memory organisations and for digital long-term preservation. Their generosity, underlying design goals, or other rea-

sons may constrain the practical applicability. The last phase of this study will cover this issue.

5.2 Survey results

From 53 distributed questionnaires we received 17 answers that could seriously be considered for analysis. So this study cannot be regarded as highly representative and comprehensive. It has to be interpreted as a first step into a deeper analysis on the transferability of methods and standards from different economically more important and dominant branches to an economic niche: digital long-term preservation, well knowing of its raising importance.

Nevertheless, we did receive important feedback from those who were simply not able to answer the questionnaire because they had not proceeded very far in establishing a digital archive. This was the case especially in one of the museums we asked, where the superior organisation, the public body in charge of the museum, has not yet recognised a preservation of the digital assets as an important issue to save cultural heritage and therefore limited the financial contribution to the basic function of the museum. Our conclusion from this feedback is, that quality management as well as long-term preservation has not reached public awareness and led to action yet. Only few stakeholders in long-term preservation have perceived the importance of standards for quality management, processes, and security for the preservation task so far.

15 out of 17 institutions were public bodies. Most (7) of those belong to a university or research institution, 5 are libraries, 4 belong to an administration, 3 are archives, and 3 data centres. We received only 2 responses from commercial institutions, although we asked 14.

Asked for the superior mission of their institution most of them identified the tasks preservation/conservation, provision and making objects accessible as key issues for their institution. From 17 institutions, 9 have defined goals and policies for their digital archive and its operation, 5 of those have even published their policies, whereas 2 institutions have no policy in place and 7 have only planned to compile a digital preservation policy.

To the question on the existence of a financial concept to the long-term provision of digital objects, 10 institutions gave a positive answer, 5 denied to have one. However, long-term in this sense corresponds to time scales between 2 years (3 institutions), 3 years (1 institution), and 5 years (5 institutions). Only one participant has a 10 year future financial concept in place.

Asked, how can the existence of the digital archive be granted after structural changes in organisation, most answers argued that this concept and question are irrelevant for public administration.

Another important response revealed, that primarily public body institutions didn't recognise an advantage for themselves, their services, and customers in being

⁵ <http://www.sei.cmu.edu/cmmi>

certified for ISO 9000 or even as trustworthy digital archive. The portability of quality management standards to the procedures and services in public administration is considered as hardly possible. Often the enormous complexity of standards is seen as main barrier to comply with them completely. Instead, standards are (mis-)used as guidelines and their principles applied to selected workflows and processes: documentation, transparency, quality control of ingested objects. An IT-concept as well as a security concept has been introduced into most of the institutions. Summarising the answers to those questions: most institutions have already thought about quality management, discussed the applicability of standards and elements derived from those standards, and follow their own interpretation of quality control and management. The study mirrored a strong demand for deeper and broader information on standards as well as support and training during the introduction of standards.

Surprisingly only 2 out of 16 institutions had appointed a quality manager.

Looking into the standards used, 12 institutions answered that they comply with standards, 3 don't. In detail it looks as follows:

ISO 9000	1 (full)
ISO/IEC 20000 ⁶	1 (full)
ITIL ⁷	3 (partially)
V-Modell ⁸	2 (mostly)
MoReq ⁹	1 (full) 1 (partially)
DOMEA ¹⁰	1 (full) 2 (mostly) 1 (partially)
DINI Certificate ¹¹	5 (full) 1 (mostly) 2 (partially)
ISO 15408 ¹²	1 (partially)
BSI ¹³ Standard 100-3	1 (partially)
BSI ¹⁴ Grundschatzkatalog	2 (full) 2 (mostly) 2 (partially)
BSI Grundschatzzertifikat	1 (partially)

One essential part of the survey was the investigation of habits regarding digital archiving systems. As we anticipated, most, 13 out of 17, institutions decided for a self-

⁶ http://www.iso.org/iso/catalogue_detail?csnumber=41332

⁷ See <http://www.itil.org>

⁸ Please refer to KBSt at <http://www.kbst.bund.de>

⁹ <http://www.moreq2.eu>

¹⁰ See KBSt: Federal Government Co-ordination and Advisory Agency

¹¹ See www.dini.de [21]

¹² See <http://www.iso15408.net>

¹³ BSI : Federal Office for Information Security

¹⁴ See <http://www.bsi.bund.de/english/topics/topics.htm>

developed software solution (only 9 documented it). This fits into the overall picture that long-term preservation is always bound to a designated community and therefore to very community specific needs. 8 out of 15 answered to use a service provider, either an external with a private contract or an administrative contract, for software development, 7 don't.

Another question looked into quality management of the service provider. Here 4 institutions answered that their service provider perform a quality management, 1 answered 'no' and 5 didn't know that. Only 1 institution mentioned ITIL as standard in use at the service provider for software development.

The type of digital objects that the interviewed institutions preserved varies from pure text formats via video and audio formats to software and interactive multimedia. In fact, there has been collected an significant amount of objects, whose only chance to survive is to be maintained in a digital preservation archive using either migration or emulation as archiving method to be available and interpretable in future.

Regarding the selection process of objects 13 participants stated to have selection criteria in place, only 3 of them published. All of them document in one or another way formal arrangements with their producers, either in form of legal regulations, frame contracts, formal license agreements or deposit contracts.

Most of the institutions (11 out of 15) have a concept in place for keeping or improving their relation to their producers.

A quality control of objects and metadata is carried out by 14 institutions, just 1 stated 'no'.

Looking into the usage aspects, most institutions know their user community and half of the institutions have already surveyed the specific demands of their user group. They use it to provide user group specific access to the digital objects. Quality can often be measured by measuring the satisfaction of the users. 6 institutions stated to measure the user satisfaction, 9 stated 'no'. Nearly one third (5) of the participants have a concept in place to continuously improve the relationship to their users.

Regarding aspects like infrastructure and security, 11 institutions stated to the question if they had defined the process and organisational structures of their institution: 11 designed, 3 specified, 5 realised, 4 published, 1 evaluated. 10 have even documented their structures, whereas 5 have no documentation.

The last two questions tested the readiness to certify themselves as trustworthy digital archive. Here we received interesting answers. Most institutions refused to answer 'yes' or 'no'. Their willingness to become a certified trustworthy digital archive strongly depends on the costs (time, effort, and money) for preparing and conducting the certification. This attitude differs from that in

different communities where e.g. an ISO 9000 certification is the basis for a successful business.

First conclusions

Summarising the first results, we regard the adoption of standards for managing quality, processes, and security as an important factor to establish trustworthy digital archives. The first results from the survey indicate that also the participants of this study, generally spoken, see the high importance of such standards for their local institutions. We also recognized severe problems in using those standards in practice. Apparently standards are applied mostly in the sense of guidelines.

The problems arising while transferring standards into new domains like long-term preservation can be traced back to the heavy complexity of those standards that affect the understanding of the standards itself in a negative way. Further reasons and potential solutions to the problem have still to be analysed in the final part of this study.

The first impression from the study leads to the finding that there is a need for a specific standard covering all relevant aspects of a trustworthy digital repository.

References

- [1] Bundesamt für Sicherheit in der Informationstechnik (2005): Common Criteria V 2.3.
- [2] ISO14721:2003 Space data and information transfer systems – Open archival information system – Reference model; see also Blue Book: <http://www.ccsds.org/docu/dscgi/ds.py/Get/File-143/650x0b1.pdf>
- [3] Digital Curation Centre und Digital Preservation Europe (2007): DCC and DPE Digital Repository Audit Method Based on Risk Assessment, V1.0.: <http://repositoryaudit.eu/download>
- [4] Erpanet Project (2003): Risk Communication Tool.: <http://www.erpanet.org/guidance/docs/ERPANETRiskTool.pdf>
- [5] McGovern, Nancy Y.; Kenney, Anne R.; Entlich, Richard; Kehoe, William R. und Buckley, Ellie (2004): Virtual Remote Control: Building a Preservation Risk Management Toolbox for Web Resources, D-Lib Magazine 10 (4): <http://www.dlib.org/dlib/april04/mcgovern/04mcgovern.html>
- [6] nestor Working Group on Trusted Repositories Certification (2006): Criteria for Trusted Digital Long-Term Preservation Repositories – Version 1 (Request for Public Comment) English Version, Frankfurt am Main.: <http://nbn-resolving.de/urn:nbn:de:0008-2006060703>
- [7] OCLC und Center for Research Libraries (2007): Trustworthy Repositories Audit and Certification: Criteria and Checklist. : <http://www.crl.edu/PDF/trac.pdf>
- [8] OCLC/RLG-NARA Task Force on Digital Repository Certification; CLR; DCC; DPE und nestor (2007): Core Requirements for digital Archives (Common Principles). : <http://www.crl.edu/content.asp?11=13&12=58&13=162&14=92>
- [9] RLG NARA Task Force on Digital Repository Certification (2005): Audit Checklist for Certifying Digital Repositories, RLG, NARA Task Force on Digital Repository Certification, Mountain View, CA.
- [10] RLG Working Group on Digital Archive Attributes (2002): Trusted Digital Repositories: Attributes and Responsibilities, RLG; OCLC, Mountain View CA.: <http://www.rlg.org/longterm/repositories.pdf>
- [11] Task Force on Archiving Digital Information (1996): Preserving Digital Information, Commission on Preservation and Access, Washington D.C.
- [12] ISO 9000:2000 Quality management systems – Fundamentals and vocabulary
- [13] ISO 9000:2005 Quality management systems – Fundamentals and vocabulary
- [14] ISO 9001:2000 Quality management systems – Requirements
- [15] ISO 19011:2002 Guidelines for quality and/or environmental management systems auditing
- [16] ISO/IEC FCD 27000 Information technology – Security techniques – Information security management systems – Overview and vocabulary
- [19] ISO/IEC 27001:2005 Information technology – Security techniques – Information security management systems – Requirements
- [20] ISO/IEC 27002:2005 Information technology – Security techniques – Code of practice for information security management
- [21] Deutsche Initiative für Netzwerkinformation: DINI Certificate Document and Publication Services 2007 [Version 2.0, September 2006]: <http://edoc.hu-berlin.de/series/dini-schriften/2006-3-en/PDF/3-en.pdf>

The Data Audit Framework: a toolkit to identify research assets and improve data management in research led institutions

Sarah Jones, Seamus Ross, Raivo Ruusalepp

Digital Curation Centre &
Humanities Advanced
Technology & Information
Institute (HATII), 11 University
Gardens, University of
Glasgow, Glasgow, G12 8QJ
s.jones@hatii.arts.gla.ac.uk

Digital Curation Centre &
Humanities Advanced
Technology & Information
Institute (HATII), 11 University
Gardens, University of
Glasgow, Glasgow, G12 8QJ
s.ross@hatii.arts.gla.ac.uk

Estonian Business Archives
Eesti Äriarhiiv Tartus,
Lembitu 6/8, 50406
Tartu, Estonia
raivo@eba.ee

At the time of writing the online toolkit was under development. It will have been tested ready for release before the iPres Conference. The Data Audit Framework will be officially launched on 1 October 2008 at the British Academy, London.

Abstract

Although vast quantities of data are being created within higher education, few institutions have formal strategies in place for curating these research outputs in the long-term. Moreover there appears to be a lack of awareness as to exactly what data are held and whether they are being managed. In response to these concerns the Joint Information Systems Committee (JISC) issued a call for proposals to develop and implement a Data Audit Framework suited to the needs of the UK higher education research communities. The Data Audit Framework (DAF) Development project was funded to produce an audit methodology, online toolkit, and a registry. Four additional implementation projects were funded to test the toolkit and promote its uptake. This paper outlines the audit methodology, introduces the online toolkit, and provides feedback on implementing the Data Audit Framework.

Overview of Data Audit Framework

Project background

One of the current challenges for UK higher education (HE) institutions is their efficient participation in the national knowledge economy. Management and reuse of research data have become critical success factors for excellence in research. While research data offer benefits they also pose risks; reaping the benefits while managing these associated risks requires knowledge of data holdings. If HE institutions are to ensure they maximise their potential to exploit and reuse research data they must be able to quickly and easily establish an overview of the data collections they hold and the policies and practices that are in place to manage them. An audit framework offers a mechanism to collect, and manage such knowledge.

The need for an audit framework was identified by Liz Lyon in the JISC-commissioned report *Dealing with*

Data: Roles, Rights, Responsibilities and Relationships. This report recommended a framework be conceived to:

enable all universities and colleges to carry out an audit of departmental data collections, awareness, policies and practice for data curation¹

The DAFD project team has produced such a framework. The methodology is simple yet flexible. As a result it can be applied across institutions irrespective of size, subject area or type of data created. A registry component will provide a mechanism to support the persistent recording of results of data audits based on DAF. This will allow organisations to share information on their data assets and curation policies while providing institutional and national perspectives to assist future data strategy development.

Project timescale

The Data Audit Framework Development project runs from April to September 2008 and is funded by the JISC under its JISC Repositories Programme.² Led by HATII at the University of Glasgow, the work is being conducted in collaboration with partners from the Estonian Business Archives, UKOLN at University of Bath, the University of Edinburgh, and King's College London. The project team has created an audit methodology and tested it in pilot audits that ran from May-July. Feedback from these audits enabled us to

¹ Lyon, L. 2007. *Dealing with Data: Roles, Rights, Responsibilities and Relationships*, p5. The recent Report of the OSI e-Infrastructure Working Group presses a similar agenda if the UK is to ensure its research institutions adapt emerging e-infrastructure realities, see: OSI e-Infrastructure Working Group. 2007. *Developing the UK's e-infrastructure for science and innovation*, www.nesc.ac.uk/documents/OSI/report.pdf

² The total value of the Grant from the JISC is £ 100,000.

refine the methodology and has yielded information that is guiding the development of the online toolkit.

A beta-version of the online toolkit will be released in September 2008 to be tested in audits at King's College and Imperial College London. Any necessary amendments will be made before the official release on 1st October 2008. The toolkit will be promoted thereafter in collaboration with the Digital Curation Centre³ and DigitalPreservationEurope (DPE)⁴. Training events are planned to assist organisations to adopt and implement the Framework. The audit toolkit will be freely available to use online or download from <http://www.data-audit.eu>. Support will also be available through the website.

The methodology and toolkit will be tested further in four JISC-funded implementation projects at University College London, King's College London, Imperial College and the University of Edinburgh. These projects should conduct some twenty audits across a range of HE departments and schools and should finish in December 2008.

The DAF Methodology

The development of the DAF methodology drew on the experiences gained by staff at HATII when developing DRAMBORA,⁵ a methodology for assessing the risks associated with digital repositories. At the outset the team recognised the value of a practice-oriented and intuitively applicable approach. DAF provides institutions with a straightforward method of collecting information on their research data assets. It has been designed so that it can be applied without dedicated or specialist staff and with limited investment of time or effort. The methodology has four stages:

1. Planning the audit;
2. Identifying and classifying data assets;
3. Assessing the management of data assets; and,
4. Reporting results and making recommendations.

The stages generate two key outputs: an inventory of data assets created during Stage 2; and a final report that incorporates recommendations on how data management could be improved. A detailed workflow of tasks and outputs within each of these stages can be seen overleaf (see Figure 1).

Audit stages

Planning the audit

There are two key objectives of the planning stage: (1) to secure organisational buy-in by establishing a robust

³ <http://www.dcc.ac.uk>

⁴ <http://www.digitalpreservationeurope.eu>

⁵ DRAMBORA: Digital Repository Audit Method Based on Risk Assessment is available at: <http://www.repositoryaudit.eu/>

business case; and, (2) to prepare as much as possible in advance of the audit so time spent on-site can be optimised. Securing agreement from top management and ensuring this commitment is filtered down is crucial. Establishing expected outcomes will assist data auditors with determining the scope and focus of the audit. By conducting background research the auditor can minimise demands placed on data creators, managers and users, and scheduling interview times and locations in advance will help ensure they are ready to contribute.

Planning of the audit involves the following tasks:

- Appoint an auditor;
- Establish a business case;
- Conduct initial research to plan the audit; and,
- Set up the audit.

Our test audits indicate that this work takes between 2-4 days, depending on the level of prior knowledge the auditor has of the department being audited and the size of the department. Where the toolkit is used internally for self-audit the initial research stages are not likely to require as much effort. The planning stage may take place over a few weeks as the auditor waits on information and responses from staff with whom interviews have been requested. During this stage a form is completed to support the capture of high level information about the organisation being audited (see DAF Methodology, Audit Form 1).

Identifying and classifying data assets

The purpose of the second stage is to establish what data assets exist and classify them according to their value to the organisation. Essentially, an inventory of data assets is compiled through a mapping exercise. The overall quality of the entire audit depends on this first knowledge-gathering exercise. Classification schemas are suggested in the inventory but will need to be tailored to the particular organisational context. The classification step will determine the scope of further audit activities, as only the vital or significant assets will be assessed in greater detail.

This stage should proceed through the following steps:

- Analyse documentary sources;
- Conduct questionnaire and/or interviews;
- Prepare data asset inventory; and,
- Approve and finalise asset classification.

Using the timing data accumulated during the test audits we can project that this work will take between 4-6 days, depending on the size and type of the organisation being audited and its data holdings. If interviews have been planned in advance during Stage 1, elapsed time should only be a couple of weeks, however this could increase if staff are unavailable to participate. During this stage an inventory of data assets, divided into groups according to their value for the organisation will be produced (see DAF Methodology, Audit Form 2)

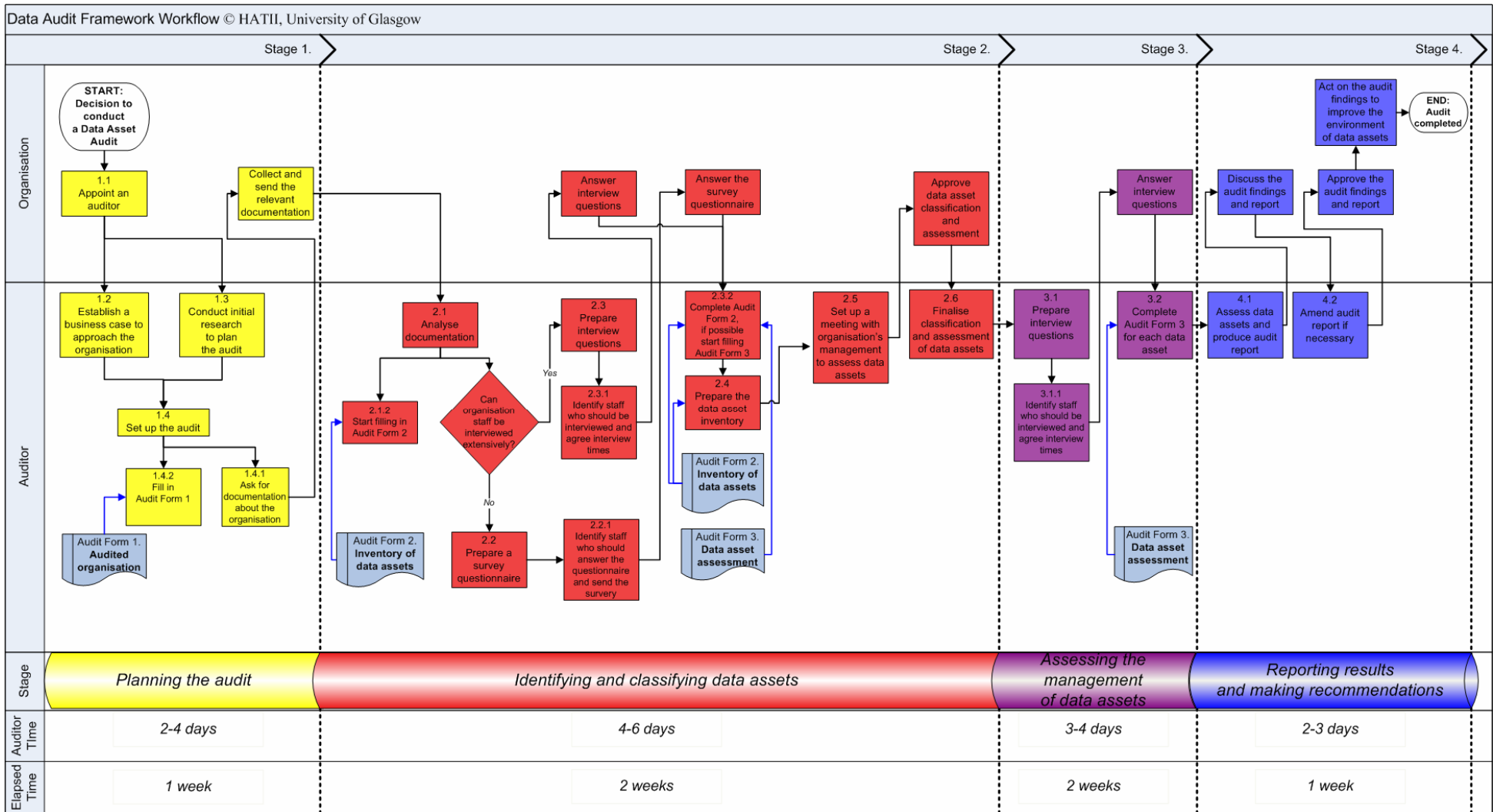


Figure 1: The Data Audit Framework Workflow

Assessing the management of data assets

The aim of this stage is to collect additional information about the data assets central to the work of the organisation. Assessing the management of these assets enables auditors assess whether the current level of resources provided is sufficient. Information collected should help identify weaknesses in data management practices and point to occasions when data are being placed at risk. During this stage several forms are completed which assist auditors in asset and context profiling (Audit Form 3A or 3B). The methodology provides two element sets to support the collection of information at different levels of detail. The level of detail adopted will be determined by the audit aims and scope set at the planning stage. Based on the pilot audits we can project that this work will take between 3-4 days, depending the number and nature of vital assets. Elapsed time is expected to be in the region of 2 weeks.

Reporting results and making recommendations

In the final stage the auditor draws together the results of the data audit to produce a final report. This report will include recommended actions to improve data management. Suggestions of relevant services and tools that could be used by the organisation to enhance their practices and services are provided in the audit toolkit and as new ones emerge we will hope to link these to the toolkit. We recommend that it would be best practice to submit the audit report to the appropriate managers within the organisation for comments before it is finalized. This stage is likely to take between 2-3 days. Elapsed time may be up to 1 week depending on the time taken to convene a meeting with management to approve the report.

Testing and updating the audit methodology

The methodology was initially tested in pilot audits based at three of the development project's partner institutions. These were split across subjects: archaeology at the University of Glasgow, engineering at the University of Bath, and GeoSciences at the University of Edinburgh. Although the audits took place in departments / schools of varying size with different data collections, the lessons learned from the pilot applications of the methodology were consistent, suggesting it is generic enough to suit diverse contexts. Moreover approaches to data curation that were encountered were consistent and confirmed the belief that auditing data assets would be of widespread benefit. We learned much from these audits and have revised the methodology as a result. We will continue to refine it as we receive further feedback from other individuals and organisations who apply it.

GUARD at the University of Glasgow

The pilot audit at Glasgow was conducted in Glasgow University Archaeological Research Division (GUARD), the archaeological research unit within the Department of Archaeology. The Unit was founded in 1989 and currently has thirty-three members of staff. It is a

commercial arm of the Department and offers a wide range of archaeological services from consultation to fieldwork and post-excavation analysis. Staff are constantly engaged in projects that result in digital data assets, such as digital images, computer aided designs, GPS/GIS, and stratigraphy and finds databases.

Implementing the methodology was straightforward. The Director of GUARD was already aware of data issues within the Unit and was keen to take part. Access was granted to the shared drives on which most data was held so much of the preparatory work and identification could be done remotely. The main challenge during the audit was arranging times to meet with staff; much of the Unit's work is conducted off-site so staff availability was poor. This was exacerbated by the audit taking place in the summer when many other staff were away on annual leave. Delays in setting up interviews increased the elapsed time. Interviews were arranged with around a quarter of the workforce. Some interviews were general discussions on data curation practices but most focused on discussion of specific data assets and were crucial in completing the assessment stage. The interviews were very useful for seeing how the Unit created and managed data and enabled the auditor to identify areas for improvement. Staff were forthcoming with suggestions of changes they felt might enhance digital curation practices within GUARD. These aspects helped feed into recommendations we could make as to how data management could be improved.

IdMRC at the University of Bath

The pilot audit at Bath was held in the Innovative Design and Manufacturing Research Centre (IdMRC). IdMRC is a research group within the Department of Mechanical Engineering. It was set up in October 2001 with funding from the Engineering and Physical Sciences Research Council's (EPSRC) IMRC programme, and is one of sixteen such centres in the UK. It has four research themes: Advanced Machining Processes and Systems (AMPS), Constraint-Based Design and Optimization (CBDO), Design Information and Knowledge (DIAK), and Metrology and Assembly Systems and Technologies (MAST). The IdMRC's work is widely supported by industry, especially from the aerospace and packaging sectors. It has emerging strengths in shoe and electronics manufacture.

No major issues were encountered when applying the Data Audit Framework in this context. An initial phone interview was held with the Director of the IdMRC to establish the scope, purpose and requirements for the audit. Preliminary research was then conducted using the Centre's website and at this stage a decision was taken as to how to compile the inventory. A snowball sampling technique was chosen, starting with interviews with the lead researchers of the four research themes. In all, ten face-to-face interviews were conducted. The interviews consisted of browsing personal and shared drives to identify assets, recording data sets in the inventory along with any additional information that could be easily captured, and discussing how the interviewee managed

the data. The resulting inventory listed 63 data sets, of which 18 were ranked as vital, 15 as important and 30 as minor. The inventory was not comprehensive but was representative of the data assets held by the Centre. Of the data assets described in the inventory, 30 were chosen for further analysis in DAF Stage 3. Much of the information required for this stage had already been collected, so there were only a few gaps and these were filled by soliciting information through e-mail queries.

GeoSciences at the University of Edinburgh

The pilot audit at Edinburgh was held in the School of GeoSciences, a leading international research centre rated 5/5* in the last Research Assessment Exercise (2001). The School hosts over 80 academics, 70 research fellows and 130 PhD students and attracts annual research grant and contract income of around £4-6 million. The School's staff contribute to one or more of five Research Groups (Earth Subsurface Science, Global Change, Human Geography, Edinburgh Earth Observatory, Centre for Environmental Change & Sustainability) and may also be involved in inter-University Research Consortia and Research Centres.

Despite the School being much larger than the other two organisations in which the methodology was applied it was still found to be appropriate. The audit began with desk research: browsing the School website, collecting annual reports and published articles, and compiling a list of research active staff including details of their research responsibilities. Interviews were conducted with thirty-five academic/research staff to compile the inventory. The interviews were semi-structured discussions during which a broad range of additional information was collected. Although this was not a comprehensive survey, the fact that the later interviews provide information duplicating that already collected indicated to the auditor that the most significant data assets had been recorded. Of the twenty-five data assets recorded only four were classified by the interviewees as vital. A detailed analysis of these assets was carried out. The audit provided crucial evidence as to the weaknesses of current approaches employed by the School to manage its data assets. The results of the audit were drawn together and a final report was produced which recommended actions for change.

Lessons learned

Several threads were raised consistently in the feedback from the pilot audits. These are categorised into five domains.

1. Ensure timing is appropriate – The initial audits were scheduled to take place in May. When planning and setting up the audits difficulties were often encountered obtaining convenient times to meet with staff. Summer holidays, exam board meetings, conferences and extended periods of fieldwork meant that the audits commenced later than anticipated. The timing of the audit should ideally coincide with the organisation's quieter period.

Originally the time suggestions given in the methodology had been in terms of person hours. As a result of their experiences applying the methodology the auditors recommended a differentiation be made between person hours and elapsed time as the lag-time between requesting information and conducting work could be quite significant. The person hours allocated for the audit were increased from 1-2 weeks to 2-3 weeks in light of the pilot audits and a suggestion was made to allow 2 months of elapsed time.

2. Plan well in advance – Setting up interviews and waiting on documentation from the organisation can take a number of weeks. To mitigate against this and avoid the audit schedule going off track, the planning stage should be started as early as possible. The person hour requirements are minimal in comparison with the likely elapsed time so planning could run concurrently with other work commitments.

3. Adopt a method suited to the context – The decision to use interviews or questionnaires will depend largely on the culture of the organisation. Where staff are known to be responsive to questionnaires, it would be worthwhile preparing and circulating one as part of the planning stage. How best to communicate with staff also depends upon organisation context and practice. One auditor found phone calls and face-to-face meetings a more effective way to engage senior management while another found personal introductions and internal advocacy a more successful approach to communicating information about the audit than email announcements.

4. Scope the work carefully – The granularity at which assets are recorded will depend on the type and quantity of data being created. The granularity could vary within the audit due to differences in types of research being conducted. Where small sets of data are created it may be most appropriate to record assets on a project or collection basis rather than individually. Convening a meeting with key stakeholders at the start of the audit to determine the scope, purpose and requirements will help focus work. The scope could be amended during the audit if necessary.

5. Collect additional information early on – Initially the audit methodology consisted of five stages, with identifying and classifying records being separate steps. All the initial audits, however, found the optimal workflow was to collect information for these stages at once. As such the original stage two and three were merged. Auditors also found it worthwhile collecting other information early in the process. Additional information was often captured when creating the inventory, for example details of file formats, software requirements, creation dates, provenance, related data assets, storage and data management. In light of these findings we have planned that the online tool will allow Audit Form 3 to be viewed when completing the inventory (Audit Form 2) so additional details can easily be entered into the relevant fields at the time of capture.

Developing the online toolkit

Background

At the time of writing the online toolkit was still under development. We have completed the system requirements stage and this has been validated.⁶ The descriptions here reflect anticipated functionality. Any discrepancies between what is planned and delivered will be noted during the tool demonstration at the iPres Conference (September 2008) and will be documented in subsequent publications about the toolkit.

Feedback from the pilots audits outlined above greatly assisted the definition of the DAF system requirements. A list of basic requirements was compiled at an update meeting and posted on the project wiki to allow additional comments to be fed back to the development team. Regular communications between the system architect responsible for defining the system requirements and authors of the methodology (one of whom had also conducted a pilot audit) ensured the appropriateness of the requirements defined.

As the online toolkit has been modelled to reflect the intentions and features of the methodology, it will facilitate planning, documentation, collection of data and final reporting. Checklists are provided and the end of each stage and contextual help will be added throughout to clarify what information is required. The main instance of the tool will be accessible over the internet at <http://www.data-audit.eu> and will be supported by secure online registries. Because we recognise that some organisations will find it unacceptable to use registries based at a second institution to store vital data about their digital assets a downloadable version will also be made available for organisations to host privately.

Functionality by audit stage

In the planning stage auditors will be guided to collect the basic information on the organisation being audited that is necessary to complete Audit Form 1. A name will be given to the audit and an upload facility will be provided for the business case. Contact details for staff within the department can be recorded and any meetings scheduled can be entered into the calendar.

In Stage 2 the auditor(s) will decide on a classification schema and set categories appropriate to the context. If a survey can be conducted the toolkit will help compile and circulate questionnaires. Alternatively the calendar system can be used to schedule interviews. Data collected at this stage will be able to be input directly into Audit Form 2. It will also be possible to enter additional data collected into Audit Form 3 ready for the next stage.

The two options for element sets in Stage 3 will be contained within separate tabs. It will be possible for the auditor to flick between one tab and another to compare the sets and make a selection as to which is most appropriate to use. Some information may already have been entered in the audit forms or pulled through from earlier stages. An additional field on both element sets will make it possible to track records by means of an automatically generated system.

The final stage of the audit requires the auditor to write a report with recommendations. Summary information and statistics will be drawn automatically from the data collected during the audit to help the auditor compile this report. The toolkit will collate information and generate a PDF appendix that contains summary details of data holdings, list of interviewees / survey respondents, and dates for the various stages of the audit. There will also be a file upload option through which the auditor may add the final audit report. It will also be possible via Stage 4 to publish audit details in the central registry. While we recognised that some organisations will not wish to have details of their data assets available in a UK-wide registry others will recognise the value of such a database to ensuring that UK higher education institutions participate in the expansion of the national knowledge economy

A status bar and calendar will be accessible throughout the audit to track progress and alert auditors to upcoming events. The toolkit will also allow files containing reports or information which helps the auditor to document the organisation, the data assets, or associated research to be uploaded. It provides 'post-it' style notes for comments to act as aide-memoirs for auditors. Each time an edit is made a new row will be added to the history table, making it possible to rollback to a previous version if necessary.

The design and implementation of the online toolkit will benefit from the experiences HATII gained constructing DRAMBORA Interactive, which was released in January 2008. The Data Audit Framework will be available to use online and the website will provide a shared area where users of the tool can seek advice and share knowledge gained from their experiences. DAF Interactive will incorporate a central audit registry into which institutions and departments will be encouraged to deposit their audit data so it can be federated at institutional and national level to assist strategy makers plan future work and to enable the HE community to improve its contribution to the UK digital economy.

⁶ Aitken, B. 2008. *The Data Audit Framework Tool: High-Level System Requirements*

Future work

The Data Audit Framework is part of a larger suite of JISC-funded data projects.⁷ The development team continues to share information and lessons learned with related projects such as the four DAF implementation studies, the UK Research Data Strategy and DataShare⁸. DAF partners are committed to collaborating across project, domain and institutional boundaries to develop tools that support data creation and management.

The methodology and online toolkit enable institutions to identify their data assets and take steps to improve data management and reuse. HATII intends to seek funding to enable it to build on the audit tool to provide additional services in the future such as a data quality assessment methodology and toolkit and a tool for assessing the 'value' of data assets. Training courses for potential auditors are being developed. Information on these and additional sources of support for institutions hoping to use the Data Audit Framework to audit their research data holdings will be provided at iPres 2008 (London) and online at <http://www.data-audit.eu>

Acknowledgements

Development of the Data Audit Framework is funded by the Joint Information Systems Committee (JISC) through a grant from its Repositories Programme. The collaboration was made possible through the Digital Curation Centre which has acted as an umbrella for this work. The authors are grateful to partners at the Universities of Bath, Edinburgh, and King's College London for piloting the DAF methodology and providing detailed feedback on its applicability. We are particularly grateful to Dr Cuna Ekmekcioglu of the University of Edinburgh and Alex Ball at UKOLN (University of Bath). We wish to thank our colleagues Brian Aitken and Matthew Barr at HATII (University of Glasgow) for developing an online version of the toolkit: Aitken for specifying the functional requirements and Barr for implementing DAF Interactive.

References

- Aitken, B. 2008. *The Data Audit Framework Tool: High-Level System Requirements*
- Jones, S., Ross, S., and Ruusalepp, R. 2008. *Data Audit Framework Methodology*, http://www.data-audit.eu/DAF_methodology.pdf
- Lyon, L. 2007. *Dealing with Data: Roles, Rights, Responsibilities and Relationships*, http://www.jisc.ac.uk/media/documents/programmes/digitalrepositories/dealing_with_data_report-final.pdf
- McHugh, A., Ross, S., Ruusalepp, R., and Hofman, H. 2007. *The Digital Repository Audit Method Based on Risk Assessment (DRAMBORA)*, <http://www.repositoryaudit.eu> ISBN: 978-1-906242-00-8
- OSI e-Infrastructure Working Group. 2007. *Developing the UK's e-infrastructure for science and innovation*, <http://www.nesc.ac.uk/documents/OSI/report.pdf>

⁷ Details of JISC's data projects are at: www.jisc.ac.uk/home/whatwedo/themes/information_environment/researchdata.aspx

⁸ For details of the UKRDS see: <http://www.ukrds.ac.uk/> and for DataShare see: <http://www.disc-uk.org/datashare.html>

Data seal of approval - assessment and review of the quality of operations for research data repositories

Dr. Henk Harmsen

Data Archiving and Networked Services
The Hague, The Netherlands
henk.harmsen@dans.knaw.nl

Introduction

Data Archiving & Networked Services (DANS) is active in the area of data infrastructure, with two main themes, namely (digital) archiving and making research data available. The field of activity of DANS covers both the social sciences and the humanities. DANS also manages its own data repository of research data.

In 2005, the founders of DANS, the Royal Netherlands Academy of Arts and Sciences (KNAW) and the Netherlands Organization for Scientific Research (NWO), gave DANS the formulation of a data seal of approval as one of its assignments. In February 2008, 17 guidelines were presented under the name Data Seal of Approval, nationally at a KNAW symposium and internationally at the first African Digital Curation Conference. This article will explain more about the backgrounds of the seal of approval: what it is and what it isn't, which international seals of approval exist, how this seal of approval matches them, what its unique selling point is, and what the plans for the future are?

What it is and what it isn't?

The data seal of approval consists of 17 guidelines that may be helpful to an archiving institution striving to become a trusted digital repository (TDR¹). The guidelines have been formulated in such a way that they are easily understandable and leave sufficient room for a broad interpretation. Standardization was not the objective as the point of departure was that the data seal of approval would remain dynamic during its first years. The seal of approval does not express any views regarding the quality of the data to be archived, but does regarding the provisions an archive has made to guarantee the safety and future usability of the data.

The seal of approval mentions 4 stakeholders: the financial sponsor, the data producer, the data consumer and the data

repository, which share an interest and are responsible for a properly functioning data infrastructure. The sponsor is advised to use the guidelines as a condition for financing of research projects. The remaining three stakeholders are addressed in the 17 guidelines. For example, the data producer is expected (three guidelines) to place its data in a TDR and to provide the research data as well as the metadata in the format requested by the data repository. The data consumer must, if it has access to or uses the information in a TDR, respect (inter)national legislation, (scientific) codes of behavior and the applicable licenses (three guidelines). The data repository, in its turn, must ensure that the archive is equipped in such a way that data producer and data consumer are able to meet their obligations. In addition, there are eleven more guidelines for the data repository, regarding organization (mission, dealing with legal regulations, quality management, long-term planning and scenarios), processes (transfer responsibility, data references, integrity and authenticity) and technical infrastructure (OAIS and automated processes).

In other words, the data repository is the stakeholder of which most is expected. Therefore, an assessment document has been formulated for the data repository which, when completed, approved and publicly published, will result in the repository being allowed to use the logo of the data seal of approval. The logo makes the repository recognizable to both data producer and data consumer.

A data repository may be able to delegate some of the guidelines to another archive that bears the logo of the data seal of approval. This way, the concerned repository does not need to execute all the guidelines in order to meet the requirements of the seal of approval.

With regard to auditing the repositories, a minimal system was chosen that is based on trust. The repository publishes its own assessment and then applies for an audit. This audit is carried out by a member of the international *DSA* (data seal of approval) assessment group² on the basis of the available assessment document. It determines whether the guidelines have been complied with and whether the logo can be awarded.

¹ The term Trusted Digital Repository (TDR) occurs in almost all seals of approval. However, it is unclear what a TDR is exactly. At the time of writing, Wikipedia does not yet have a description of the concept. Main point of such a repository is 'trust'. It is the basis of the data seal of approval.

² The international *DSA* assessment group will be launched in the fall of 2008

International initiatives

The text accompanying the seventeen guidelines states³ that these 'are in accordance with, and match national and international guidelines regarding digital data archiving'. In this section, I will explore the mentioned initiatives in slightly more detail.

*Catalogue of Criteria for Trusted Digital Repositories*⁴ - NESSTOR

This catalogue has identified criteria that can help in the evaluation of the reliability of digital archives at both the organizational and the technical level. The criteria were defined in close cooperation with a broad range of data institutions and information producers. One of the objectives is to offer a tool enabling archiving institutions to archive and demonstrate reliability. The catalogue is also an opportunity for arriving at the certification of repositories, with a 'standardized national or international process'. Again, 'reliability' or 'trust' plays a role here. The catalogue can be used for conceiving, working out and eventually implementing a 'trusted digital long-term repository' and for working out (in various stages) of a self-assessment.

The criteria catalog employs over fifty criteria organized into fourteen sections that are arranged into three areas of attention namely: Organizational framework, Object management, and Infrastructure and Security.

*Digital Repository Audit Method Based on Risk Assessment (DRAMBORA)*⁵ of the Digital Curation Centre (DCC) and DigitalPreservationEurope (DPE)

The DRAMBORA toolkit is available to support internal audits of archiving institutions. To this end, the party responsible for the archive has the challenge of tracking down the weaknesses, while at the same time acknowledging the strength of the archive.

DRAMBORA helps track down the many risks any archiving institution runs. This takes place in the form of process description:

- A detailed description of the organization (mission, and activities);
- Formulation of possible risks, organizational as well as technical, that may occur;
- Evaluation of the impact of these risks and making them manageable and controllable.

DRAMBORA gives support by means of templates for the description of risks and codes to assess the severity of the risks. Apart from that, it is an open process which must be shaped by the party responsible for the repository. There is, however, a list of examples of possible risks.

³ Data Seal of Approval, chapter 0.3 Guidelines. See:

<<http://www.datasealofapproval.org>>

⁴ See: <<http://edoc.hu-berlin.de/docviews/abstract.php?id=27249>> [site visited 15 August 2008].

⁵ See:

<<http://www.digitalpreservationeurope.eu/announcements/drambora/>> [site visited 15 August 2008].

The philosophy of the DRAMBORA authors is clear: by monitoring closely what people are doing and how they are doing it, a repository is capable of keeping the risks involved in archiving of data under control.

Further, the Research Library Group (RLG)⁶ developed the *Trustworthy Repositories Audit & Certification (TRAC): Criteria and Checklist*.

This criteria checklist comprises three sections, arranged into a various aspects, in their turn subdivided into more than eighty criteria.

The paper *Foundations of Modern Language Resource Archives* of the Max Planck institution in Nijmegen⁷ must not remain unmentioned. The document describes a data seal of approval specifically for language bodies. A language resource archive (LRA) must meet nine principles.

The Research Information Network in the UK⁸ developed the *Stewardship of digital research data: a framework of principles and guidelines*. This document is built up of 5 principles, spread across 40 guidelines.

The German Initiative for Network Information (DINI) developed the *Certificate Document and Publication Services of the Deutsche Initiative für Netzwerkinformation*⁹, a certificate mainly intended for institutional repositories with their own Document and Publication Services.

Synthesis

The guidelines of the data seal of approval can be seen as a basic set of the above proposals. The data seal of approval wants to facilitate 'awareness' at the archiving institutions. It can serve as a first step toward a 'heavier' assessment and certification. The authors see the data seal of approval as supporting for example TRACK and DRAMBORA. The objective of the data seal of approval was mainly to try and convince archiving institutions to start paying attention to quality management.

Unique selling point

The data seal of approval (DSA) as developed by DANS has a number of unique features: The DSA is oriented toward scientific data, not primarily toward publications. The DSA not only pays attention to the archiving institution, but also to the data producer and the data

⁶ See:

<<http://www.crl.edu/content.asp?l1=13&l2=58&l3=162&l4=91>> [site visited 15 August 2008].

⁷ Peter Wittenburg, Daan Broeder, Wolfgang Klein, Stephen Levinson, of the Max-Planck-Institute for Psycholinguistics in Nijmegen, The Netherlands, and Laurent Romary of the Max Planck Digital Library in Munich, Germany. See: <<http://www.lat-mpi.eu/papers/papers-2006/general-archive-paper-v4.pdf>> [site visited 15 August 2008].

⁸ See: <<http://www.rin.ac.uk>> [site visited 15 August 2008].

⁹ See: <<http://edoc.hu-berlin.de/series/dini-schriften/2006-3-en/PDF/3-en.pdf>> [site visited 15 August 2008].

consumer. This encourages the idea of shared responsibility.

As indicated before, the DSA is not in conflict with for example TRAC, but is rather a step toward it. Where TRAC chooses standardization, the DSA opts for 'trust'. This way of working does on the other hand match the custom of peer review in the scientific world.

The DSA also focuses on smaller organizations. The DSA is relatively light and therefore easy to implement. Openness, dynamics and speed are possible in the actual implementation.

The DSA is formulated as points of attention, not as solutions. Finally, the DSA offers possibilities for subcontracting archiving and still meet the requirements of the DSA. This will be appreciated by research groups with their own data projects.

Future

In 2009, DANS will comply with the data seal of approval and its policy is aimed toward being on the way to meeting the TRAC criteria. Furthermore, DANS uses the code for information security¹⁰.

DANS strives toward internationalization of the data seal of approval. The previously mentioned DSA assessment group will be launched in the fall of 2008, and that same year, four pilots will be planned in The Netherlands as a first step in the area of certification of the DSA.

¹⁰ CVI - *The Code voor Informatiebeveiliging* is the Dutch version of the British Standards 7799, which was later published as ISO/IEC 17799 as international standard for information security in organizations. It is a general code applicable to all institutions that work with information.

Updating DAITSS - Transitioning to a web service architecture

Randall Fischer, Carol Chou, Franco Lazzarino

Florida Center for Library Automation
5830 NW 39th Avenue
Gainesville, FL 32605, USA
rf@ufl.edu, cchou@ufl.edu, flaz@ufl.edu

Abstract

The Florida Digital Archive (FDA) is a long-term preservation repository for the use of the libraries of the public universities of Florida. The FDA uses locally-developed software called DAITSS, which was designed to perform the major functions of Ingest, Archival Storage, Data Management and Dissemination in the OAIS reference model. A DAITSS 2 project is in process to re-write the application based on a distributed, Web services model. This paper describes the major changes in store for DAITSS 2.0, the rationale behind them, and the issues involved in their design and implementation. These changes include: moving from a monolithic to distributed processing environment; implementation of modular RESTful services; incorporation of existing tools, services, and registries; and revising the internal data model to be more conformant with the PREMIS data.

Introduction

The Florida Digital Archive (FDA) is a long-term preservation repository for the use of the libraries of the public universities of Florida. It has been in operation since late 2005, and as of July 1, 2008 has archived 52,000 information packages comprising 3.6 million files (10.4TB). Nine universities have agreements with the FDA to archive their submissions, which are being ingested at an average rate of 30-60 GB per day.

The FDA uses locally-developed software called DAITSS, which was designed to perform the major functions of Ingest, Archival Storage, Data Management and Dissemination in the OAIS reference model. DAITSS implements format-specific preservation strategies including normalization, migration and localization. ([Caplan 2007])

DAITSS was a pioneering digital preservation system. When it was designed and developed, there were few models of true preservation repositories and few external tools available for performing specific functions such as format validation and metadata extraction. It is somewhat remarkable that in three years of FDA operations, no major functional flaws have been discovered and few enhancements to functionality are pressing. The architecture of the application, however, requires major redesign. DAITSS was coded as a monolithic, self-contained system. A DAITSS 2 project is in process to re-write the entire system based on a distributed, Web services model.

The fundamental principles governing the original design of DAITSS have not changed. These include:

- strict conformance to the OAIS functional model;
- a requirement that the archived data store be self-defining, so that if the DAITSS system were lost, all known information about archived objects could be recovered from the data store itself;
- data once written to archival storage cannot be altered; modified objects are in effect new objects;
- original versions of archived files must be retained unaltered.

In conformance with these principles, files are modified only during the Ingest process as the SIP is transformed into the AIP. DAITSS relies upon format normalization and migration as preservation strategies, and these are implemented as part of Ingest. All files in the SIP as originally submitted are retained unaltered in perpetuity, but other versions may be derived and added to the AIP.

The basic unit of storage and processing is an Information Package. Each Information Package consists of an XML descriptor and all of the content files required to assemble one (and possibly more) representations of an information object. The Information Package is the only unit of input and output; that is, even if only a single file in an AIP is needed, the entire IP must be disseminated.

Because many years may pass between the time a file is ingested and when it requires some preservation treatment, dissemination requests are filled by a three-step process. In the first step, the AIP is exported from the repository and placed in the Ingest queue as a SIP. In the second step, the AIP-cum-SIP is re-ingested, and undergoes file identification, validation, and transformation processing according to the current version of the software. In the final step, the resulting AIP is reformatted into a DIP and delivered to the requestor.

This model will be retained in DAITSS 2. It has worked well in practice and in fact has beneficial side-effects. For example, the ingest model makes updates extremely simple, and the dissemination model allows the FDA to implement migration on request or mass migration depending on the circumstances.

Another governing principle was to use standard formats and metadata schemes whenever possible. However, at the time DAITSS was initially developed, there were few

applicable standards to choose from. METS is used as the format for SIP, AIP and DIP descriptors, and within the METS document standard schema are used for format-specific technical metadata for the few formats for which such schema exist. These include the Audio Engineering Society's draft AES schema for audio, the Metadata for Images in XML schema (MIX) for raster images, and the TextMD schema maintained by the Library of Congress for text. The Preservation Metadata: Implementation Strategies (PREMIS) Working Group was meeting as a committee while DAITSS 1.0 was being coded, but the PREMIS Data Dictionary had not yet been issued, so DAITSS 1 is only partially PREMIS compliant.

Design goals for DAITSS 2

Papers While time has shown the principles, approach and basic functionality of DAITSS to be sound, the current generation of software has a number of problems:

- The application was in some respects over-built, anticipating problems and functional requirements which never materialized. Unnecessary logical complexity makes the software difficult to maintain and configure.
- DAITSS is written as monolithic Java application, hindering its ability to scale. Simple functions such as virus checking take a significant portion of processing time, but cannot easily be offloaded to an independent server.
- There is a high degree of coupling between components, making it hard to extend and enhance the application. Adding support for a new format, for example, requires changes to dozens of classes, database schema, and XML schema.

The second generation of DAITSS will address these flaws. It will also improve PREMIS compliance throughout, by bringing the internal data model into closer conformance with the PREMIS three-part Object model (file, representation, bit-stream), and by making extensive use of PREMIS Object and Event descriptions.

Eliminate unnecessary complexity

Two features, initially thought to be desirable, have proved problematic. The first is the concept of preservation levels. DAITSS depositors (called "affiliates") are allowed to associate each file format with any of three preservation levels to be applied to files contained in their SIPs: BIT, FULL or NONE. NONE specifies that files of a given format will not be archived at all. BIT specifies that files of a given format will be archived but not subject to format transformation. FULL indicates that files will be normalized and/or migrated as appropriate.

Although it seemed like good customer service to give FDA affiliates these options, in practice it has been confusing to affiliates and problematic for the archive. The option NONE was intended to allow an affiliate to assemble a single package for multiple purposes; for example, for archiving and for loading into a digital asset management system. An unexpected problem is that

files in formats that cannot be correctly identified because of DAITSS limitations might be assigned preservation level NONE and dropped from the AIP. In DAITSS 2 we will assume that if a file is in a SIP it is intended for archiving, and affiliates will be responsible for assembling appropriate SIPs.

The distinction between BIT and FULL has also proved difficult to sustain, and there seems to be little added functionality in maintaining it. Since DAITSS always retains files from the SIP as originally submitted, if an affiliate wants to ignore a migrated version they can always do so. DAITSS 2 will eliminate the entire concept of preservation level and attempt full preservation treatment for all files.

The second issue involves "global" files and a kind of transformation called "localization." Global files are sets of files included in many packages. Commonly these are files needed to validate XML descriptors, such as DTDs and schema. Rather than storing them redundantly in thousands of AIPs, the global files are stored once in separate packages and referenced, as necessary, by links from other AIPs. Although this seemed like a good idea at the time, the maintenance of global files has added considerable complexity to the code. Analysis shows that the space savings are only about 1.6% of the archive store. DAITSS 2 will eliminate the concept of global files, and will include all required files in each AIP.

Localization is a DAITSS 1 function where a reference within an archived file to an external file (for example, a schema) is rewritten to refer to a locally archived version. This requires DAITSS to keep both the original and localized versions of the file. DAITSS 2 will skip localization at the file level, and instead modify validators to dynamically resolve references to the external file from a local cache.

Break up the beast

Two features, initially thought to be desirable, have proved problematic. DAITSS 2 will be comprised of simple, independent components that each perform one simple function. It is a requirement that each component can be tested and developed independently of any other component. This will make it simpler to modify or extend existing functions and to integrate new functions. For example, it would be possible to add a new risk assessment service to the current chain of processing without modifying any other service. Dividing DAITSS into separate components will also allow us to parallelize time-consuming tasks such as virus checking and checksum calculation.

Further, we believe that exposing each functional component as a stand alone service will allow researchers to extend the system into novel workflows.

In short, rather than providing major changes in functionality, we wish to simplify and support existing functions but with a wider scope.

Implementation

This The second generation of the DAITSS software will take a Web services approach. There are two main competing architectural styles for Web services today: a

Remote Procedure Call (RPC) style, and the Representational State Transfer (REST) style detailed by Roy Fielding ([Fielding 2000]). SOAP is an example of the RPC style, while REST is the basis for the classic view of HTTP used on the Web. The Web service APIs provided by Google, Amazon and Yahoo typically offer both styles of access to their services. However, the REST APIs are significantly more popular: Amazon has reported that REST style requests comprise 80% of their web service traffic ([Anderson 2006]).

Experience has shown that SOAP applications exhibit a high degree of coupling between services. This state of affairs results from very application-specific SOAP actions that must communicate data structures from one service, to the client, to other web servers. This has led to ever expanding sets of standards and complex frameworks to support what was, initially, a Simple Object Access Protocol.

In contrast, the REST approach is centered around resources. In HTTP, the most successful example of a RESTful architecture, there are only six operations and each of them are atomic. PUT creates a named resource, GET retrieves it, POST modifies it, and DELETE removes it. HEAD retrieves simple metadata for the resource.

The state of a RESTful application is maintained as a set of external resources. A client program effects the progress of the application by performing incremental changes using defined operations on externally stored resources. Such limitations allow, counter-intuitively, far greater flexibility on the part of client-based applications, illustrating the key design strategy in software engineering of using the least powerful language to accomplish a task ([W3C 2006]).

In its purest form, the state of an application is driven by resources that contain links to other services, the so-called Hypertext As The Engine Of Application State (HATEOAS). In DAITSS this is illustrated by the Action Plan service described below. Briefly, this service is given data identifying and characterizing a format, and returns the location of an appropriate transformation service that can effect format migration and normalization. The archival policy of the FDA is thus driven by a very simple service which publishes links to other services.

The DAITSS Storage Service

Rather than implementing a wholly new version of DAITSS at some time in the future, our plan is to gradually morph DAITSS 1 into DAITSS 2 by pulling out pieces of the code and replacing them with newly written Web services that perform the same function. Our first Web service has already been incorporated into the production version of DAITSS used by the Florida Digital Archive: a simple storage service loosely based on the Amazon S3 Web service. The implementation of this storage service resulted in a significant performance increase for the FDA.

Each AIP is assigned an intellectual entity identifier (IEID) and its constituent files and descriptors packaged together as a GNU tar file. The MD5 checksum of the tar file is computed as well as the checksums of the

individual files it contains. The assembled package is then submitted to two geographically isolated servers using SAN-attached file systems as long term storage. The package-level checksum is used to ensure that the initial transmission completed successfully, and is also retained for subsequent fixity checking on the stored AIP.

A typical HTTP conversation for the initial store is shown for an AIP that has been assigned the IEID E20080715_AAACAZ; the client stores the AIP using the HTTP PUT function.

Request:

```
PUT /sil003/E20080715_AAACAZ HTTP/1.1
User-Agent: DAITSS v1.5
Host: storage.fcla.edu:3000
Content-MD5: 2thsYe6iN5MvIBAJ5UMWCQ==
Content-Type: application/octet-stream
Content-Length: 32044941
[ ... inline data ... ]
```

Response:

```
HTTP/1.1 201 Created
Connection: close
Date: Mon, 11 Aug 2008 16:08:42 GMT
Content-Length: 0
```

Possible success and error conditions with the associated response status codes include:

Success

201 The resource was created

Client Error

400 Missing resource name in PUT request

403 Duplicate package name

405 Storage location is full

409 Checksum error

411 Invalid request headers

Server Error

500 Specific server error message included

The DAITSS 2 Service Architecture

We next describe the entirely services-based architecture planned for the second generation of DAITSS. The current monolithic application will be decomposed into a set of relatively simple Web services, some of which are described below. The composition of each service into a complete Ingest process will require preservation events to be recorded as they occur, and later assembled into a complete record of the archiving process. Therefore each function will create an event description expressed in PREMIS XML, which will ultimately be assembled into the AIP descriptor. Main components of the Ingest Process are shown in Figure 1.

The Description Service

File format identification and validation is a central function of DAITSS. In DAITSS 2, each data file is sent to the Description Service for identification, validation and characterization. The service uses DROID for a preliminary identification of the file format, which is used to select the appropriate validator. If DROID returns the information that the file is identified as multiple formats associated with different validators, the

most appropriate validator is selected by the service. For the formats most commonly presented to the Florida Digital Archive, a modified version of JHOVE is used as the validator, and the preliminary format is used to select the initial JHOVE validation module. JHOVE may include in its output the information that the file is actually described by multiple formats; if so the most appropriate format is selected by the service. The result of JHOVE validation and characterization is then parsed and mapped into PREMIS, and the JHOVE format information is converted back to a PRONOM format identifier.

A PREMIS XML document for that file is returned by the Description Service to guide further Ingest processing. The returned PREMIS document has three sections: an object section that includes a single PRONOM format identifier and technical metadata according to an extension schema appropriate to that format; an event section that describes the outcome of the validation, including any anomalies found; and an agent section that identifies the service used. An abbreviated version of an example document is shown below.

```
<object xsi:type="file">
  <objectIdentifier>
    <objectIdentifierType>
      DAITSS2</objectIdentifierType>
    <objectIdentifierValue>
      E20080715_AAACAZ/florida.tif
    </objectIdentifierValue>
  </objectIdentifier>
  <objectCharacteristics>
    <compositionLevel>0</compositionLevel>
    <size>3001452</size>
    <format>
      <formatDesignation>
        <formatName>TIFF</formatName>
        <formatVersion>4.0</formatVersion>
      </formatDesignation>
      <formatRegistry>
        <formatRegistryName>
          PRONOM</formatRegistryName>
        <formatRegistryKey>fmt/8
        </formatRegistryKey>
      </formatRegistry>
    </format>
    <objectCharacteristicsExtension>
      <mix:mix xmlns:xsi="http://www.w3.org/
        2001/XMLSchema-instance" >
        ...
      </mix:mix>
    </objectCharacteristicsExtension>
  </objectCharacteristics>
</object>
<event>
  <eventIdentifier>
    <eventIdentifierType>DAITSS2
  </eventIdentifierType>
  <eventIdentifierValue>1</eventIdentifierValue>
</eventIdentifier>
  <eventType>Format Description</eventType>
  <eventDateTime>2008-07-17T12:32:50
</eventDateTime>
```

```
<eventOutcomeInformation>
  <eventOutcome>Well-Formed and valid
</eventOutcome>
  <eventOutcomeDetail>
    <eventOutcomeDetailExtension/>
  </eventOutcomeDetail>
</eventOutcomeInformation>
</event>
<agent>
  <agentIdentifier>
    <agentIdentifierType>uri</agentIdentifierType>
    <agentIdentifierValue>
      http://daitss.fcla.edu/describe
    </agentIdentifierValue>
  </agentIdentifier>
  <agentName>Format Description Service
</agentName>
  <agentType>Web Service</agentType>
</agent>
```

The Action Plan Service

The Action Plan Service is sent the PREMIS document produced by the Description Service and returns a simple XML document containing one or more links to services to be used to transform (migrate or normalize) the associated file. If DAITSS is not capable of transforming a given format, or if a particular file contains too many anomalies to be reliably transformed, the document will contain, instead of links, a stanza noting the limitation.

The Action Plan service succinctly specifies the migration and normalization policy of an installation of DAITSS. The service illustrates a key feature of the RESTful approach, which is to let links drive the process of ingest. An example of a document returned by the action plan service follows.

```
<instructions>
  <normalization>
    <transformation>
      http://daitss.fcla.edu/transform/wave_norm
    </transformation>
  </normalization>
  <migration>
    <limitation>codec not supported
  </limitation>
  </migration>
</instructions>
```

The Action Plan Service is driven by a set of XML documents that serve a dual function: they are used internally to specify the transformation services to be applied, and they are published externally to document our archival policy:

```
<action-plan format="WAVE" date="2008-07-02"
author="Andrea Goethals, FCLA">
  <processing>
    <normalization>Each audio stream in the WAVE
      file will be normalized into an uncompressed
      PCM(LPCM) audio stream with sample size of 16
      bits/sample.
    <transformation>
      http://daitss.fcla.edu/transform/wave_norm
    </transformation>
```

```

    <limitations>
      <supported-codec>PCM</supported-codec>
      <supported-codec>MP3</supported-codec>
    </limitations>
  </normalization>
</processing>
<strategy>
  <original>Migrate to newer WAVE versions or to
  an open, standardized and well supported audio file
  format that is to be a good successor to WAVE.
  </original>
  <normalized> Migrate to an open, standardized and
  well supported audio stream format that is losslessly
  compressed.
  </normalized>
</strategy>
<timetable>
  <item action="review" date="2009-07-02"/>
  <item action="revise" date="2009-07-02"/>
  <item action="short-term" date="2009-07-02">
    Write or locate a converter which converts WAVE
    files with data in one of audio encoding formats
    listed in 3.1 to WAVE files in LPCM format.
  </item>
</timetable>
</action-plan>

```

The Transformation Service

The current version of DAITSS provides both normalization and migration of data files. The second generation of DAITSS will support these transformations via a collection of Transformation Services. A file is submitted to the appropriate Transformation Service as specified by the Action Plan service; the transformed file is returned via HTTP. It is possible for multiple files to be produced as output from a single submission. For instance, DAITSS may normalize a PDF file into a collection of TIFFs, one per page. For cases like these, the Transformation Service returns a composite document using the MIME multipart/mixed standard. In some cases the Transformation Service is a locally developed program. In many cases a Transformation Service is simply an HTTP wrapper around an external, probably open source, program such as Ghostscript, ffmpeg, mencoder, or libquicktime. For these cases, a simple specification of the action of the program suffices to build the service.

```

<transformations>
  <transformation ID='WAVE_NORM'>
    <instruction> ffmpeg -i #INPUT_FILE# -sameq -a
    codec pcm_s16le #OUTPUT_FILE#
    </instruction>
    <extension>.wav</extension>
    <software>FFmpeg version SVN-r9102
    </software>
    <configuration> --prefix=/opt/local--
    prefix=/opt/local --disable-vhook--
    mandir=/opt/local/share/man --enable-shared --
    enable-pthreads --disable-mmx
    </configuration>
    <dependency>libavutil version: 49.4.0

```

```

    libavcodec version: 51.40.4
    libavformat version: 51.12.1
  </dependency>
</transformation>
  <transformation ID='AVI_NORM'>
    <instruction>mencoder #INPUT_FILE# -oac pcm -
    ovc lavc -lavcopts vcodec=mjpeg --o
    #OUTPUT_FILE#
    </instruction>
    <extension>.avi</extension>
  </transformation>
  <transformation ID='MOV_NORM'>
    <instruction>lqt_transcode -ac rawaudio -vc mjpg
    #INPUT_FILE# #OUTPUT_FILE#
    </instruction>
    <extension>.mov</extension>
  </transformation>
  <transformation ID='PDF_NORM'>
    <instruction>gs -sDEVICE=tiff12nc
    -sOutputFile=#OUTPUT_FILE# -r150 -dBATCH
    -dNOPAUSE #INPUT_FILE#
    </instruction>
    <extension>page%d.tif</extension>
  </transformation>
  ....
</transformations>

```

The AIP Service and subsequent processing

At this point both the original file and any derived versions are submitted to an AIP Service, which acts as a holding area for this intellectual entity. The PREMIS object and event descriptions are also saved. When the last file in the SIP has been fully processed, a complete AIP descriptor is assembled combining information from the original SIP descriptor with the saved object and event information. Finally, the entire package is sent to the Storage Service, which, as noted above, distributes the AIP to multiple locations.

Conclusion

As noted above, we believe that dividing complex services into simple, well understood components will allow the creation of novel preservation workflows. One new function under consideration is a risk assessment service, which will accept information extracted from an AIP descriptor and return the preservation risk associated with the packages.

However, the architecture has other advantages. For one thing, it will make it possible for the Florida Digital Archive to share services with other preservation repositories. Several institutions and projects are developing Web services based systems or components, including (but not limited to) The National Archives (UK), the California Digital Library, PLANETS and PRESERV. The FDA (and other DAITSS users) will be technically capable of integrating externally-written services if rights and organizational issues allow.

In addition, while the first generation of DAITSS is actively maintained and distributed as open source software, we have made little effort to promote its use in the community, as our experience has been that DAITSS

is overly complex to configure and difficult to maintain. The Florida Center for Library Automation has neither the resources nor the mandate to exert significant effort supporting external sites. We expect that DAITSS 2 will be much easier to configure and operate, and that other institutions would find it attractive to implement the system or some of its component services. The architecture is particularly advantageous to local sites, which could customize the distribution version of DAITSS by supplying their own action plans and services as needed.

References

- Anderson, T. 2006. *WS-* vs the REST*. The Register April 29, 2006.
http://www.theregister.co.uk/2006/04/29/oreilly_amazon
- Caplan, P. 2007. *The Florida Digital Archive and DAITSS: A Working Preservation Repository Based on Format Migration*. International Journal on Digital Libraries, 20 March 2007, doi: 10.1007/s00799-007-0009-6. Available at <http://www.springerlink.com> and http://www.fcla.edu/digitalArchive/pdfs/IJDL_article.pdf
- Fielding, R. 2000. *Architectural Styles and the Design of Network-based Software Architectures*. University of California, Irvine
<http://www.ics.uci.edu/~fielding/pubs/dissertation/top.htm>
- W3C, 23 February 2006. *The Rule of Least Power*. W3C Technical Architecture Group.
<http://www.w3.org/2001/tag/doc/leastPower.html>

Conceptual Framework for the Use of the Service-oriented Architecture-Approach in the Digital Preservation

Christian Saul, Fanny Klett

Fraunhofer Institute Digital Media Technology
Business Area Data Representation & Interfaces
Ehrenbergstrasse 31
98693 Ilmenau, Germany
{Christian.Saul; Fanny.Klett}@idmt.fraunhofer.de

Abstract

This paper presents a conceptual framework for the use of the SOA¹-approach in the digital preservation. The focus of this work reflects the service composition part within the SOA service concept. Previously released approaches have been separately using process-oriented models to describe the behaviour of services, and structure composition models to represent service interactions. In this paper, the authors attempt to combine the even mentioned disjunctive models to obtain a comprehensive model, which represents both, the structure and the behaviour of the services. For this purpose, the novel SCA²-BPEL³ serves as basis for the implementation in a future-oriented SOA-compliant digital preservation software system. The SCA-model specifies the architecture of the intended system, while the BPEL-model indicates the behavior of each service, which is defined in the SCA-model. We can conclude that the SCA-BPEL approach is well-suited for building a scalable, adaptable and service-oriented software system. The knowledge gained from the conceptual framework will serve as a basis for future digital preservation developments.

Keywords

Service oriented Architectures, Digital Preservation, Service Component Architecture, Business Process Execution Language

Introduction

Digital preservation refers to the management of digital information over time. It is defined as a long-term, error-free storage of digital information, in terms of retrieval and interpretation, of the entire time span the information is required for. Long-term is defined as "long enough to be concerned with the impacts of changing technologies, including support for new media and data formats, or with a changing user community. [1]

Due to the rapid grow of digital information the data transfer volume to digital repositories rises continuously. This makes it necessary to find new levels of automation in digital archiving and preservation solutions. The increasing diversity in size and complexity of new digital resources implies that the repository systems must become highly automated and adaptable to various types of input, storage, access, and simultaneously to the users. The level of automation and technology support in current digital preservation solutions is low, and involves several manual stages. The scalability of existing preservation solutions has been poorly demonstrated by now, and solutions often have not been properly tested against diverse digital resources, or in heterogeneous environments.

Research in digital preservation domain has moved away from trying to find one ideal solution to the digital preservation problem towards focusing on the definition of practical solutions for preservation situations. This approach has to utilize the experts' know-how in memory institutions, to implement industry standards, and moreover, to involve solutions that are scalable and adaptable to heterogeneous environments.

Related Work

SOA is an emerging approach [2] that describes flexible software architectures, which can offer proper solutions to the above-mentioned problems of digital preservation systems.

The functionality of those architectures is provided as loosely coupled services over standardized interfaces. The aim is to map the business processes through a suitable composition of various services, in order to achieve a high flexibility related to process variability.

[3] associates four key concepts (see figure 1) with a service-oriented architecture namely application front-end, service, service repository, and service bus. The focus of the SOA-approach is the service concept itself.

Services are packaged software resources, which are well-defined, self-contained modules that provide standard business functionality and are independent of the state or

¹ Service-Oriented Architecture (SOA)

² Service Component Architecture (SCA)

³ Business Process Execution Language (BPEL)

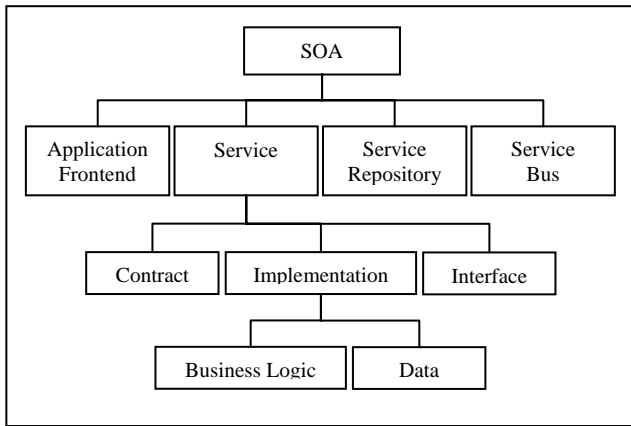


Figure 1: Elements of SOA [3]

context of other services. They are described in a standard language, have a published interface, and communicate with each other by requesting application of their operations, in order to collectively support a common business task or process. [4]

[5] states the service concept in one concise sentence that summarizes the important facts. Services are autonomous, platform-independent entities that can be described, published, discovered, and loosely coupled in novel ways.

Building a software system typically requires combining multiple existing services. These composite services can be recursively composed with other services into higher level solutions. According to [6], the two models of service composition in SOA's are both, the process-oriented- and the structural composition model described below.

Process-Oriented Model

The process-oriented composition combines services by using a workflow model to define a new service component. BPEL [7] is the most applied specification for this composition model.

Business Process Execution Language (BPEL). BPEL defines a model and a grammar for describing the behavior of a business process based on interactions between the process and its partners. The interaction with each partner occurs through Web Service interfaces, and the structure of the relationship at the interface level is encapsulated in a partnerLink. The BPEL process defines how multiple service interactions with these partners are coordinated to achieve a business goal, as well as the state and the logic that are necessary for this coordination. BPEL also introduces systematic mechanisms for dealing with business exceptions and processing faults. Moreover, BPEL introduces a mechanism to define how individual or composite activities within a unit of work are to be compensated in cases where exceptions occur or a partner requests reversal. [7]

Summarizing it can be stated that BPEL is concerned with business logic and the sequence of operations, which are performed to execute an individual business process.

Structural Composition Model

In contrast to the process-oriented composition, structural composition focuses on identifying the participating components, and the component connections that represent component interaction. The SCA [8] is the specification of a structural composition model for SOA.

Service Component Architecture (SCA). SCA represents a flexible SOA architecture standard for building composite applications using reusable services and extends, and complements prior approaches to implementing services. The SCA builds on open standards such as Web services. The SCA is based on the idea that business function is provided as a series of services, which are assembled together to create solutions that serve a particular business need. [8]

The SCA is concerned with what components exist in a business application, what services those components offer, what services reference those components, how the components are connected together, what endpoint addresses and communication methods are used for the connection, what policies are applied to components and to the connections between them.

Service Component Architecture Assembly Model. The SCA Assembly Model (see figure 2) [9] consists of a series of artefacts, which define the configuration of an SCA domain in terms of composites, which contain assemblies of service components, the connections and related artefacts, which describe how they are linked together.

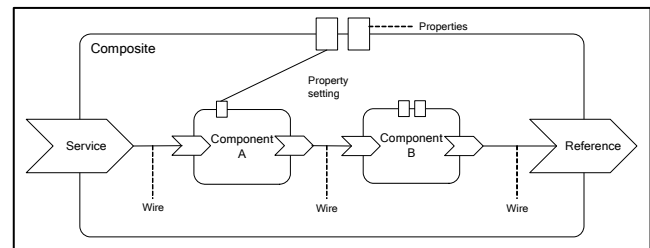


Figure 2: Service Component Architecture Assembly Model

One basic artefact of SCA is the component, which represents the unit for the construction of the SCA. A component consists of a configured instance of an implementation, which provides business functions. The business function is offered to be used by other components as a service. Implementations may depend on services provided by other components. These dependencies are called references. Implementations can provide properties, which are data values, which influence the operation of the business function. The component configures the implementation by providing values for the properties and by wiring the references to services provided by other components.

The SCA describes the content and linkage of an application in assemblies called composites. Composites can contain components, services, references, property declarations, and wires, which describe the connections between these elements. Composites can group and link components built from different implementation technologies by allowing appropriate technologies to be used for each business task. In turn, composites can be used as complete component implementations: providing services, depending on references, and with settable property values. Such composite implementations can be used in components within other composites by allowing for a hierarchical construction of business solutions, where high-level services are implemented internally by a set of lower-level services. The content of composites can also be used as a group of elements, which can contribute to build higher-level compositions.

Combination of Process-Oriented and Structural Composition Model

[10] argues that the implementation of the components as BPEL processes within an overall SCA assembly represents a good combination. Our work focuses on the application of the novel SCA-BPEL service composition approach to build a SOA-compliant digital preservation system. We combine the above explained service composition models to get a combined view of the structure and sequence: The SCA shows the structure of our composite service application while BPEL processes determine the flow sequence for each operation.

SOA-Approach in the Digital Preservation

The underlying Digital Preservation Workflow

The aim of our project is to develop a SOA-compliant digital preservation system by including a methodology that facilitates preservation work based on Web-Services. The intended digital preservation system refers to a preservation workflow that starts with a Pre-Ingest Phase over a Transfer Phase to an Ingest Phase.

Pre-Ingest is the preparatory phase for transfer of records from producer to the repository. During this phase the producer describes and normalises the content to comply with the requirements of the repository.

Transfer is the phase where the storage of records is transferred from the producer to the repository and between the repositories. It involves the transfer agreement, an optional test transfer, the actual transfer of records and their metadata, validation of the records, and acceptance from the repository.

Ingest is the phase where the repository is checking the transferred records, normalizes the transferred records and prepares them for long-term preservation in its storage, and for metadata management.

Here, we would like to point out that the OAIS workflow [1] starts with an Ingest Phase. In other words,

we add a number of stages prior to the general workflow to extend the functionality of the indented system, and consequently to enhance the system support in a more extensive manner. Figure 3 illustrates this fact.

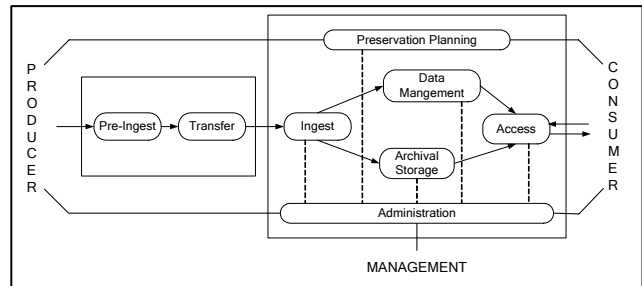


Figure 3: Extended OAIS Workflow Model

In the following the Pre-Ingest Phase is pointed out as an example to illustrate the whole process.

Building Components

To develop an innovative product with no obvious precedent, an understanding of the users and their capabilities, their current tasks and goals, the context of use of the product, and the constraints on the products performance is required. In order to communicate the user needs, requirements, objectives and expectations have to be discussed, refined, specified, and probably re-scoped.

Therefore, a variety of data gathering methods to collect sufficient, relevant and appropriate data is needed so that a set of stable requirements can be produced.

The most important needs arise from the data gathering methods that are focusing on the Pre-Ingest Phase:

- Creation of records management classification scheme
- Automate the process of appraisal
- Routine technical transactions (conversion into archival file formats, etc.)
- Compare documents and access restrictions against requirements from archival institution
- Analyze records (metadata, duplicates, classification)

Upon gathering the user needs the next step is to assemble appropriate components, which address the adequate needs. Each component has a dedicated task to fulfill functional requirements stated by user needs.

The following components for the Pre-Ingest Phase result:

- Technical Identification Component
- Digital Repository Requirements Component
- Metadata Improvement Component
- Migration Service Component

- SIP¹ Generating Component

The task of the Technical Identification Component is to identify the technical characteristics of the digital records like file formats, or the accompanying metadata formats. The Digital Repository Requirements Component analyses the digital repositories in terms of the respective requirements for long-term preservation. This could be mandatory file formats, or specific metadata elements. In order to conform to the requirements, some file format changes are necessary. Hence, the Digital Repository Requirements Component additionally delivers a list, or proposes the tools to transform the digital records into a digital repository compliant format. The Metadata Improvement Component uses the provided information of the Digital Repository Requirements Component, and the technical characteristics of the Technical Identification Component to improve the metadata related to the records. In the same way, the Migration Service Component uses the output of the Digital Repository Requirements Component and the Digital Repository Requirements Component to migrate the digital records according to the requirements of the intended digital repository. The SIP Generating Component prepares the digital records and their metadata according to the SIP configuration accepted by the digital repository.

Modeling the Software Architecture

Based on the even identified services the architecture of the intended system can be constructed. The next step is to build the structure of the system through composing the services among each other. The result of this step is an architecture, which can supply information about their components: what services they offer, what services they use, how they are linked together, etc.

To model the structure of the SOA-compliant digital preservation system we have chosen for the new SCA-approach, because the SCA extends and complements prior approaches to implementing services, and it builds on open standard such as Web services. The SCA provides a model, both for the composition of services, and for the creation of service components by including the reuse of existing application function within the SCA composites.

As mentioned above, in the following we picked out the Pre-Ingest Phase in our digital preservation workflow to reveal the SOA-approach. The other phases can be implemented in the same manner.

The following figure 4 illustrates our Pre-Ingest SCA composite assembled from a series of components. The Pre-Ingest composite consists of five components as defined above, one offered service and three references to external services. The five components offer both references and services, and they are connected by wires, which describe the connections between those elements.

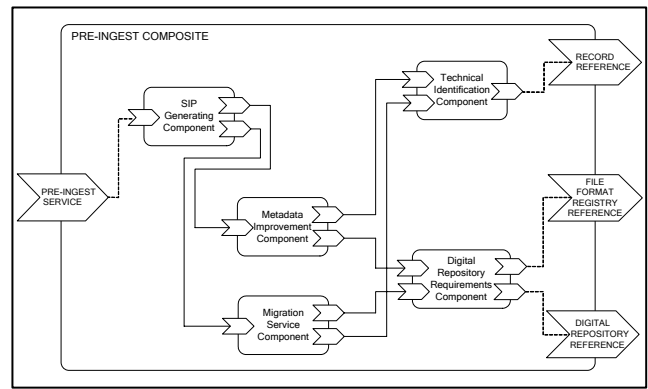


Figure 4: Pre-Ingest SCA Composite

The architecture of the intended digital preservation system only represents the fixed structure of the system without any information about the implementation of the components. BPEL aims at addressing this problem in particular, and its role is reflected in the following section.

Implementing the Software Architecture

The BPEL is a language for specifying business process behaviour based on Web Services. The processes in the BPEL export and import functionality by using Web Service interfaces exclusively, and determine the flow sequences for individual operations.

In the following the Digital Repository Requirements Component is picked out to highlight the possible business logic implementation with BPEL. The Digital Repository Requirements Component is aimed at gathering requirements of digital repositories in terms of file formats, mandatory metadata elements, standards, etc., and at informing about appropriate transform tools.

Figure 5 visualizes the implementation of the Digital Repository Requirements Component with the BPEL.

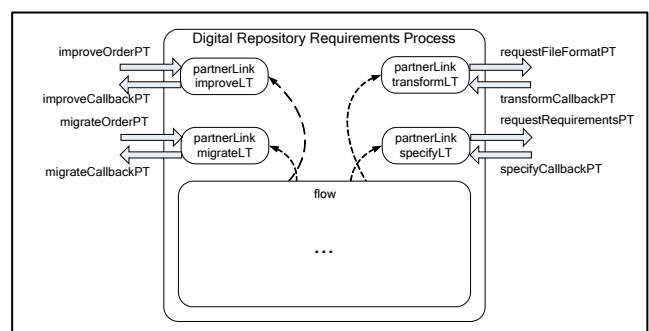


Figure 5: Digital Repository Requirements Component BPEL Implementation

¹ Submission Information Package (SIP)

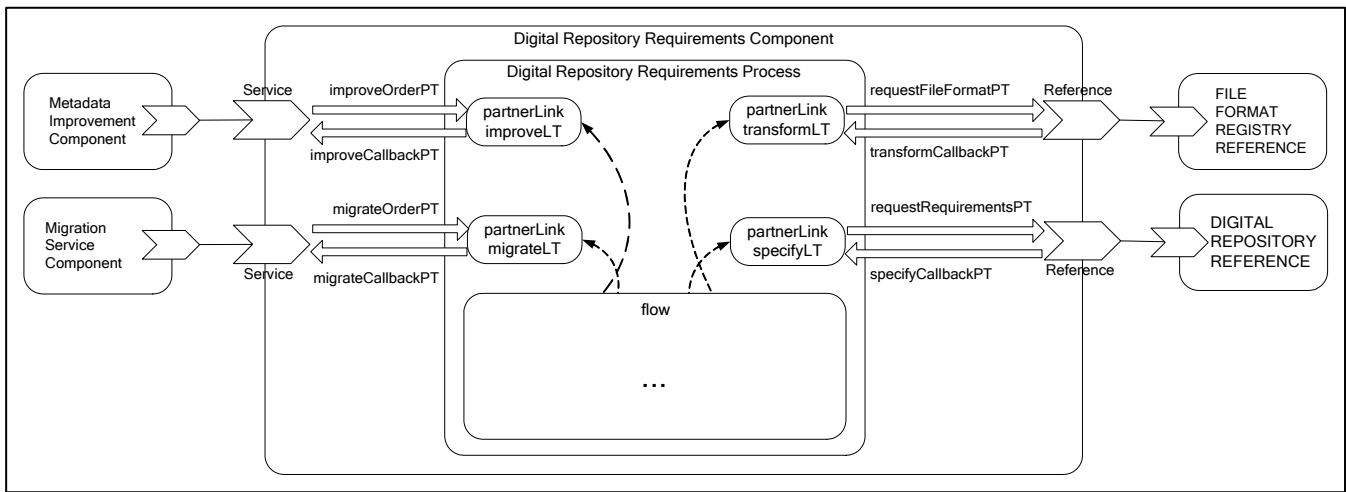


Figure 6: Pre-Ingest SCA Composite with a BPEL Implementation

The BPEL process invokes two services, one to request tools for the transformation of file formats, and another one to request the requirements of the indented digital repository for long-term preservation. In addition, the process offers two services one to provide recommendations for improvement, and the other to provide recommendations for migration. These relationships are captured in four partnerLinks.

The BPEL process definition appears as the implementation of the prior defined Digital Repository Requirements Component. The process definition is the foundation for a deployment by a BPEL engine. A BPEL engine interprets and executes business processes described in the BPEL.

Against this background, we have two independent models to describe a SOA-compliant digital preservation software system. The SCA-model reflects the architecture of the intended system. It specifies the interaction of services, and assembles them together to form a composite application. In contrast to the SCA-model, the BPEL-model specifies the behaviour of each service, which is prior defined in the SCA-model. It is obvious that a combination of the explained models is sufficient.

Modeling and Implementing the Software Architecture

The SCA and the BPEL are complementary technologies. The BPEL is an execution language while a SCA captures only the dependencies. But a BPEL process can be an implementation type of a service within the SCA.

BPEL captures relationship between the process and an interactive web service as a partnerLink with different roles linked to port types. The SCA maps the partnerLink with a single role (port type) to a reference.

Figure 6 shows the SCA+BPEL Pre-Ingest Composite consisting of the components, references and services which are linked together. In our case the partnerLinks define two roles, one for the BPEL process, and one for the partner. Depending on the message flow direction, one of them becomes a reference, and the other one becomes a service. In our solution, two BPEL process interfaces are exposed as a service entry point of the composite while two other partnerLinks are mapped as references. The BPEL process does not know the implementation, the references, and their binding. This loose coupling and flexibility is a power of the SCA architecture.

Conclusion

This paper has presented a novel SCA-BPEL service composition approach applied to build a conceptual framework towards a SOA-compliant scalable and adaptable digital preservation system.

The example showed the Pre-Ingest phase of the Digital Preservation workflow according to the OAIS model. Each stage in our workflow has been modeled according to the SCA-BPEL approach. The SCA describes the structure of a workflow component (i.e. Pre-Ingest Phase, Transfer Phase, Ingest Phase, etc.), and the connections between them. The sequences, in which the particular services are involved, are determined by the BPEL. It implements the business logic of the digital preservation system.

Directions for Future Work

The knowledge gained from the conceptual framework will serve as a basis for future digital preservation developments after completing a proper formative and summative evaluation in several iterative stages of the system design.

Acknowledgements

This work is partly supported by the Grant EU project N° 216746 PReservation Organizations using Tools in AGenT Environments (PROTAGE), FP7 and the Fraunhofer Institute for Digital Media Technology, Ilmenau.

References

- [1] Consultative Committee for Space Data Systems: “*Reference Model for an Open Archival Information System (OAIS)*”, Washington, DC, CCSDS Secretariat, 2002.
- [2] Mike P. Papazoglou; Willem-Jan von den Heuvel: “*Service oriented architectures: aproaches, technologies and research Issues*“, The VLDB journal, volume 16, 2007.
- [3] Dirk Krafzig, Karl Banke, Dirk Slama: “*Enterprise SOA: Service Oriented Architecture Best Practices*”, Prentice Hall International, 2005.
- [4] Fremantle, P., Weerawarana, S., Khalaf, R.: “*Enterprise Service*”, Commun. ACM 45(10), 2002.
- [5] Papazoglou, M.P., Traverso, P., Dustdar, S. Leymann, F.: “*Service-oriented Computing: State of the Art and Research Challenges*”, in Computer – Innovative Technology for Computer Professionals - IEEE Computer Society, 11/2007.
- [6] Mike Edwards: “*Relationship between SCA and BPEL*”, Open SOA Collaboration, 2007.
- [7] OASIS: “*Web Services Business Process Execution Language Version 2.0*”, <http://docs.oasis-open.org/wsbpel/2.0/OS/wsbpel-v2.0-OS.html>, 2007.
- [8] OASIS Open SOA: “*Service Component Architecture (SCA)*”, <http://oasis-opencsa.org/sca>, 2007.
- [9] OASIS Open SOA: “*Service Component Architecture Assembly Model Version 1.00*”, http://www.osoa.org/download/attachments/35/SCA_AssemblyModel_V100.pdf, 2007.
- [10] Francisco Curbera: “*Component Contracts in Service-oriented architectures*”, in Computer – Innovative Technology for Computer Professionals - IEEE Computer Society, 11/2007.

RODA and Crib

A Service-Oriented Digital Repository

José Carlos Ramalho
DI/UM - CCTC
jcr@di.uminho.pt

Miguel Ferreira
DSI/UM
mferreira@dsi.uminho.pt

Luis Faria
D GARQ
lfaria@iantt.pt

Rui Castro
D GARQ
rcaastro@iantt.pt

Francisco Barbedo
D GARQ
frbarbedo@iantt.pt

Luis Corujo
D GARQ
lcorujo@iantt.pt

Abstract

In 2006 the Portuguese National Archives (Directorate-General of the Portuguese Archives) engaged in the development of an OAIS compatible digital repository system for long-term preservation of digital material. Simultaneously, at the University of Minho a project called CRiB was being devised which aimed at the development of a wholesome set of services to aid digital preservation. Among those services were format converters, quality-assessment tools, preservation planning and automatic metadata production for retaining representations' authenticity. This paper provides a detailed description of both projects and discusses how these may be integrated into a complete digital preservation solution based on currently available archiving and preservation standards, e.g. OAIS, EAD, PREMIS, METS and ANSI/NISO Z39.87.

Introduction

In mid 2006, the Portuguese National Archives (Directorate-General of the Portuguese Archives) have launched a project called RODA (Repository of Authentic Digital Objects) aiming at identifying and bringing together all the necessary technology, human resources and political support to carry out long-term preservation of digital materials produced by the Portuguese public administration.

As part of the original goals of the RODA project was the development of a digital repository capable of ingesting, managing and providing access to the various types of digital objects produced by national public institutions. The development of such repository was to be supported by open-source technologies and should, as much as possible, be based on existing standards such as the Open Archival Information System (OAIS) (SYSTEMS 2002), METS (of Congress 2006), EAD (of Congress 2002) and PREMIS (Group 2005).

At a higher level the OAIS model is composed by three mega processes (ingest, administration and dissemination). In RODA we have specified the workflows for each one of those. Ingest process takes care of new information packages additions to the repository (Submission Information Packages - SIP): the SIP structure was formal specified.

During ingest SIP are transformed into AIP (Archival Information Package): we had to specify a data model for storing AIPs. Dissemination process takes care of consumer requests delivering information packages to them (DIP - Dissemination Information Packages). We had to specify one or more DIP structures for each type of Digital Object stored in RODA repository. Currently RODA is capable of storing and give access to the following types of Digital Objects: Text Documents, Still Images and Relational Databases.

Normalization plays an important role in RODA. It was not possible to archive every kind of text document or every kind of still image. Even with databases, each Database Management System has its own datamodel. So we had to take measures towards format normalization. Every Digital Object being stored in RODA suffers a normalization process: Text Documents are normalized into PDF; Still Images are normalized into uncompressed TIF; Relational Databases are normalized into DBML (Ramalho et al. 2007) (Database Markup Language).

The RODA project is divided in different components, being the base component the Fedora Commons framework. Fedora implements the common digital repository features, as digital object (and metadata) storage abstraction and relationships between objects, and it can be extended by the Fedora's Generic Lucene Search engine. On top of that, the RODA Core Services implements all the base RODA services, which can be accessed programatically. Finally, the RODA Web User Interface allows the end user to easily browse, search, access and administrate all the digital objects, metadata and ingest, preservation and dissemination tasks.

In spite of all the efforts invested in the development of RODA, there was still no support for real active digital preservation. Once the materials got into the archival storage they remained untouched and, therefore, susceptible to technological obsolescence, especially at the format level.

At the same time, at the University of Minho, a project called CRiB (Conversion and Recommendation of Digital Object Formats) was being devised. This project aimed at assisting cultural heritage institutions as well as consumers in the implementation of migration-based preservation interventions. Among those services were format converters, quality-assessment tools, preservation planning and automatic metadata production for retaining representations'

authenticity.

The CRiB system was developed as a Service Oriented Architecture (SOA) and is capable of providing the following set of services:

- File format identification;
- Recommendation of optimal migration options taking into consideration the individual preservation requirements of each client institution;
- Conversion of digital objects from their original formats to more up-to-date encodings;
- Quality control assessment on the overall migration process - data-loss, performance and format suitability for long-term preservation;
- Generation of preservation metadata in PREMIS format to adequately document the preservation intervention and retain the objects' authenticity.

After obtaining supplementary funding to continue the development of RODA, the team decided to use CRiB as its preservation planning and execution unit.

The RODA project follows a service-oriented architecture to facilitate the parallel development and update and allow heterogeneous technology and platform independence between the various components. The CRiB project is also service-oriented, to allow the implementation of services that are only possible in specific platforms and technologies. This paper provides a description of both projects and about the integration of CRiB as a RODA component, allowing the use of its features in the ingest normalization and metadata generation tasks, on the preservation planning and events, and even on the dissemination services.

RODA project

Digital archives are complex structures usually composed of human resources, high-end technologies, policies and information. The RODA project set ground for a series of studies on all these axes. Its original goals were rather ambitious, namely:

- to define the functional requisites for a digital archive, its consumers and compliant applications;
- to devise conceptual, logical and content models for a digital archive;
- to identify the set of metadata schemas that are necessary to support all functions of the digital archive (descriptive, technical, structural and preservation metadata);
- to identify technical and organizational requisites;
- to develop a digital repository system capable of storing and preserving digital objects for the amount of time defined in the law;
- to develop software modules that integrate with available records management software applications;
- to develop an acquisition and ingest policy for the digital objects produced by Portuguese public institutions;
- to devise a preservation plan and policy for the digital archive;

- to promote a study on business models capable of financing the digital archive;
- to define taxonomies of significant properties for each class of digital objects to be supported by the archive in order to implement quality control mechanisms.

One of the stages of this project consisted in the development of a repository system capable of preserving digital information and making sure that that information remained accessible to its potential consumers without ever compromising its authenticity. This repository would serve as a basis for the development of a fully functional digital archive capable of ingesting and managing large quantities of digital objects at the national level. In its first version, the repository was expected to handle a small range of object classes, namely, text documents, raster images and relational databases.

Architecture

RODA follows the Open Archival Information System Reference Model (OAIS) (SYSTEMS 2002). OAIS identifies the main functional components that should be present in an archival system capable performing long-term preservation of digital materials. The proposed model is composed of four principal functional units: Ingest, Data management, Archival storage and Access; and two additional units called Preservation planning and Administration. Figure 1 depicts how these functional units interact with each other and with all the stakeholders of the repository (internal and external).

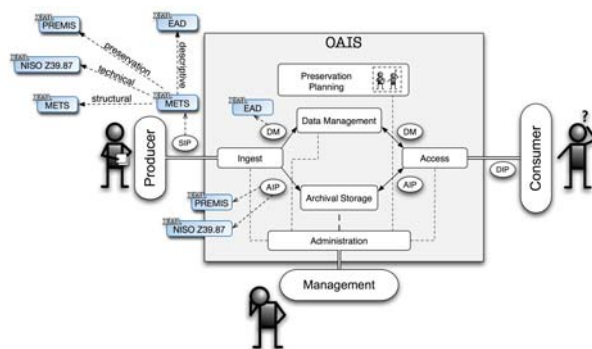


Figure 1: RODA general architecture

Before engaging in any technical developments, a collection of functional requisites was assembled by RODA's archival team (Barbedo 2006) and a study on currently available repository platforms was conducted. In this study, DSpace (COMPANY and LIBRARIES) and Fedora (Lagoze et al. 2005) were compared against this collection of requisites.

DSpace outperforms Fedora on most of the requisites. Nevertheless, the project team ended up choosing Fedora as its development platform. Even though DSpace, as it comes out of the box, combines a broader range of ready-to-use features and user-friendly interfaces, it lacks flexibility and expansibility. One very pragmatic example of this is the support metadata schemas other than Dublin Core. One would

have to go through a tremendous amount of work to make DSpace compatible with more complex descriptive metadata structures such as EAD (of Congress 2002).

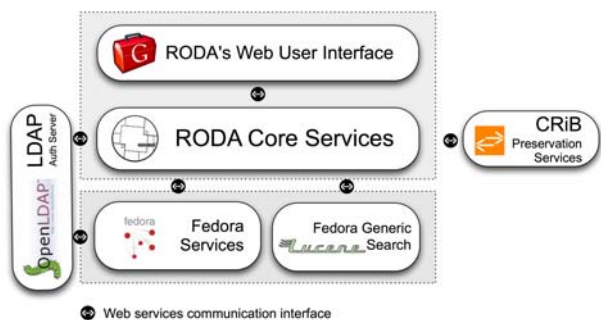


Figure 2: RODA service oriented architecture

Figure 2 depicts the overall architecture of services that compose RODA's repository. On the bottom one may find the basic services provided by Fedora. These account for elementary tasks at the Data Management and Archival Storage level. Examples of such services are ingest, add a data stream to an object, get data stream, purge object, find objects and list data streams. For a complete list of the services provided by Fedora (Project). Fedora search capability is supported by Apache Lucene and its authentication procedures go through a LDAP server (Lightweight Directory Access Protocol).

RODA Core Services are responsible for carrying out more complex tasks such as implementing the complete set of actions that compose the ingest workflow, querying the repository in more advanced and abstract ways and carrying out administrative functions on the repository. The same LDAP server previously described is used by RODA's Core Services for authenticating repository users.

On top of the RODA's Core Services lays the RODA's Web User Interface (RODA-WUI). This layer handles all the aspects of the graphic user interface for producers, consumers, archivists, system administrators and preservation experts. The RODA-WUI components are supported by the Google Web Toolkit and all communication is done via AJAX and Web services technologies.

Ingest process As previously described, Fedora only provides a set of very basic services that developers are expected to extend in order to create a fully working repository system. This includes the development of graphical user interfaces and the characterization of most of the OAIS functional units outlined at the beginning of this section. The ingest process was the first of the units to be developed.

The ingest process is responsible for accommodating new materials into the repository and takes care of every task necessary to adequately describe, index and store those materials. For example, in this stage the repository may transform submitted representations to normalized formats adequate for long-term preservation and request the user to add descriptive metadata to those objects to facilitate their future retrieval using available search mechanisms. It is also

common practice to store the original bit-streams of ingested materials together with the normalized version (just in case a more advanced preservation strategy comes along to rescue those old bits of information).

New entries come in packages called Submission Information Packages (SIP). When the ingest process terminates, SIPs are transformed into Archival Information Packages (AIP), i.e. the actual packages that will be kept in the repository. Associated with the AIP is the structural, technical and preservation metadata, as they are essential for carrying out preservation activities.

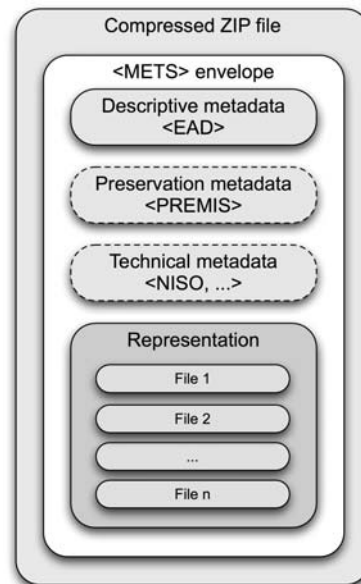


Figure 3: Submission Information Package structure

The SIP is the format used to transfer new content from the producer to the repository. It is composed of one or more digital representations and all of the associated metadata, packaged inside a METS envelope. The structure of a SIP supported by RODA is depicted in Figure 3. The RODA SIP is basically a compressed ZIP file containing a METS document, the set of files that compose the submitted representations and a series of metadata records. Within the SIP there should be at least one record of descriptive metadata in EAD-Component format¹. However, one may also find preservation and technical metadata inside a submission package, although this last set of metadata is not mandatory as it is seldom created by producers. Nevertheless, it was felt important that RODA should support those additional SIP elements for special situations such as repository succession, i.e. when ingested items belong to another repository that is to be deactivated.

¹An EAD record does not describe a single representation. In fact, EAD is used to describe an entire collection of representations. Our SIP includes only a segment of EAD, sufficient to describe one representation, i.e. a $\langle c \rangle$ element and all its sub-elements. The team has called this subset of the EAD an EAD-Component.

Before SIPs can be fully incorporated into the repository they are submitted to a series of tests to assess its integrity, completeness and conformity to the ingest policy.

If any of the validation steps fails, the SIP is rejected and a report is sent to the archivists group as well as to the producer. The producer may then fix the problem and resubmit a new version of the SIP.

Access interface The access component establishes an interface between the archive and the end user (i.e. the consumer). This functional unit is able to locate an AIP by querying the data management and retrieve it from the archival storage unit. The AIP is then transformed to a Dissemination Information Package (DIP) and delivered to the consumer.

Data model

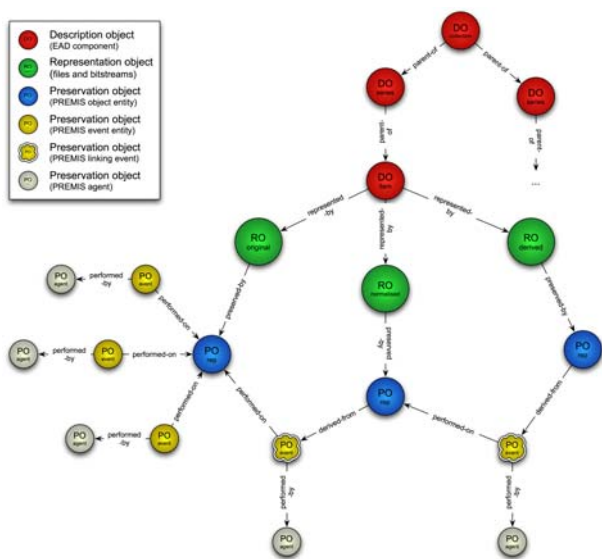


Figure 4: RODA Data model

RODA's data model is atomistic and very much PREMIS-oriented (Figure 4). Each intellectual entity is described by an EAD-component metadata record (DO nodes in Figure 4). These records are organized hierarchically in order to constitute a full archival description of a collection but are kept separately within the Fedora Commons content model. Relationships between EAD-components are created using Fedora's own RDF linking mechanism.

Additionally, each leaf record of a hierarchical collection (i.e. a file or an item) is linked to a representation object (RO nodes in the figure), i.e. a fedora object that embeds all the files and bit-streams that actually compose the digital representation. Finally, each of these objects are linked together by a set of PREMIS entities that maintain information about the digital object's provenance and history of events (PO nodes).

Each preservation event that takes place in the repository is recorded as a new preservation-event node (i.e. PO

event nodes in the figure). Special events, like format migrations, establish relationships between two preservation-representation nodes. These are called linking events in this context. Each preservation event is executed by an agent, whether this be a system user or an automatically triggered software application. The agent that triggered the event is recorded in PO agent nodes.

CRiB

CRiB² is a project being developed at the University of Minho that delivers a set of preservation services intended to aid client institutions in the planning and execution of migration-based preservation interventions (Ferreira, Baptista, and Ramalho 2007; 2006). Preservation planning is supported by a recommendation service that makes educated decisions on the best migration options available and takes into consideration the individual requirements of each client institution. The preservation execution component is handled by a large set of migration services that may be composed together to create more complex migration paths. To better understand how the system works, one should describe each of its constituent components.

Architecture

Figure 5 illustrates CRiB general architecture. The application layer illustrates how client applications may take advantage of the services provided by the CRiB. Examples of such applications may be custom programmes developed by individual users or complex applications such as digital repository systems like DSpace, Fedora, Eprints or RODA.

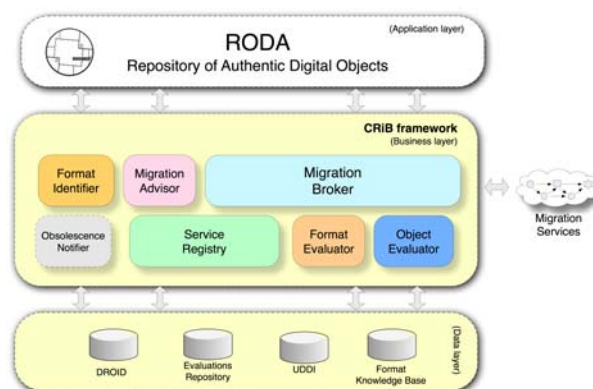


Figure 5: CRiB architecture

The middle layer illustrates the set of components that actually constitute the CRiB.

The Format Identifier, as the name suggests, is a service capable of identifying the underlying encoding of a digital representation. Client applications responsible for preserving digital objects must be able to identify, characterise and validate the integrity of its objects, if possible without human intervention. This service is indispensable in accom-

²Conversion and Recommendation of Digital Object Formats

plishing this goal. Furthermore, it enables format descriptions to be uniform across all components of the CRiB - format descriptions belong to a controlled vocabulary defined by the PRONOM file format registry developed by the National Archives of the UK (Darlington 2003).

The Obsolescence Notifier is responsible for monitoring the level of disuse of recognised file formats. When a given file format is at risk of becoming obsolete (e.g. when a new version of a format is published), this component will make sure that the adequate preservation events are triggered. CRiB does not actually implement this component as it has largely been the focus of an Australian initiative called AONS (Curtis et al. 2007).

The CRiB also delivers a large set of migration services for converting still-images and text-documents between various formats. The Service Registry component is responsible for storing information about these services. This allows the CRiB to rapidly discover which migration services are available and ready to be used. The metadata elements used within this component are based on the Universal Description, Discovery and Integration (UDDI) standard.

The Migration Broker is an additional component that is responsible for making sure that composite migrations are carried out atomically from the CRiB's point-of-view. Additionally, the broker is in charge of measuring the performance of each migration service for purpose of assuring quality control. Performance is measured according to multiple criteria, such as: availability, stability, throughput, cost, size of the outcome representation and the number of outcome files in the resulting representation in relation to the original one. The results of these evaluations are then stored in an additional component called the Evaluations Repository. This repository is used by the Migration Advisor to determine the most apt migration services available.

The Object Evaluator is the component accountable for detecting the data-loss that might occur during the conversion process. This assessment is fundamental in determining the success of a migration process and to adequately document the preservation intervention. This component works by comparing the representation submitted to migration with its converted counterparts. Evaluations are performed according to fixed, but extendable, set of criteria, usually known as significant properties. These constitute the set of attributes that are expected to be maintained intact during the preservation intervention. They constitute the range of attributes that characterize the digital representation as a unique intellectual entity, independently of the format in which the representation is encoded.

The evaluations performed by the Object Evaluator are returned to the client application for documentation purposes and stored in the Evaluations Repository (again, to aid the subsequent recommendation process).

The evaluation report sent to the client follows the structure of the Event entity described in the PREMIS Data Dictionary. This entity includes elements for describing the type of event (e.g. Migration), the date and time of occurrence, the agent that carried out the event and detailed information regarding the outcome of the event (e.g. the amount of changes on significant properties that occurred during the

migration process).

The Format Evaluator provides information about the current status of file formats. This information enables the Migration Advisor to determine to which formats are more adequate for long-term preservation by looking at its technical characteristics. The Format Evaluator works by questioning the Format Knowledge Base, i.e. a data store of known facts about digital formats. In the future this service could be replaced by other sources of information such as services provided by PRONOM or other external services such as Google Trends.

The Migration Advisor is in charge of preservation planning. It accomplishes this by generating suggestions of migration alternatives and works by confronting the preservation requirements outlined by client applications and its users with all the accumulated knowledge about the quality/performance of each individual migration service (or composition of services). It is important to point out that this component learns from each executed migration. During a migration, the system records its quality/performance in terms of data loss, status of involved formats and migration performance. Using this information, the Migration Advisor is able to rank all the available migration options and produce an appropriate suggestion for a migration intervention. Migration suggestions typically include the target format and the access point(s) of the most optimal migration services and/or migration paths. Additional information on the inner workings of this recommendation process may be found on (Ferreira, Baptista, and Ramalho 2007; 2006).

RODA meets CRiB

As previously described, CRiB offers a large set of preservation services that may be used by any client institution, application or individual user in order to maintain their collections of digital objects in interpretable and in up-to-date encodings making sure that the risk of losing important representational features is kept to a minimum.

Most of the services delivered by CRiB are relevant to RODA. The following scenario depicts how these services may be used semi-automatically by the repository:

During the course of its activity, RODA is expected to ingest and archive a large set of digital objects, most of these being submitted by its own producers and very likely, encoded in various formats. After ingesting these objects, RODA typically invokes the file Format Identification service provided by CRiB as to determine whether or not the recently deposited digital object is well formed and recognised as being in one of the preservation formats stated in the preservation policy of the archive.

If the ingested object is not already in an acceptable preservation format, the CRiB may be queried to find available migration services capable of carrying out the correspondent normalisation. After obtaining a list of possible migration services, RODA may opt to invoke one of the suggested access points in order to obtain a novel representation of that object.

Together with the new representation, CRiB returns a migration report that thoroughly describes the outcome of the preservation action, especially in what concerns the effects of the intervention on the significant properties of the original object. This report may then be stored by RODA as preservation metadata to fully document the undertaken intervention. Preservation metadata serves the purpose of providing evidence on all preservation actions applied to any given object in the repository and is considered a fundamental tool in the preservation of authentic digital objects.

The repository also makes use of CRiB's migration services to create derivative representations of its preserved objects with the goal of making them more adequate for dissemination.

Routinely, the repository consults the Obsolescence Notifier to check if any of its preservation formats is at risk of becoming exceptionally outdated. If so, the repository may request CRiB's Migration Advisor to provide a recommendation for a new preservation format and engage in the migration of all of its outdated objects.

Preservation management within RODA is handled by a scheduler in which a special user, i.e. the preservation expert, may define the set of rules that trigger specific preservation actions. Preservation actions comply with a common API, so creating and installing new actions in the repository is as easy as copying the programme file to the right directory on the server. These actions may invoke remote services such as the ones provided by CRiB, but must be deployed locally for the sake of conformity. The locally deployed actions must handle all remote service invocations and handle all possible exceptions that might occur.

The scheduler allows the preservation expert to configure the rules that will select relevant objects for a particular preservation intervention as well as scheduling the intervention itself.

Conclusion and Future Work

As previously described, RODA has been planned to be a complete digital repository providing functionality for all the main units that compose the OAIS reference model. RODA fully implements an Ingest workflow that not only validates SIPs, but also takes care of the whole negotiation process between the archive and the producers of information. RODA also accounts for Access providing different ways to search and navigate over available metadata as well as visualizing/downloading stored digital objects. Administration components were also developed allowing archivists to change the descriptive metadata and define rules for preservation interventions such as scheduling integrity checks on all stored digital objects, initiate the migration process of certain representation class/formats, or control which users or groups are authorized to perform certain actions within the repository.

Although RODA covers most of the functional components described in the OAIS reference model, there was still

one very important one missing in its design ? Preservation Planning.

According to OAIS, Preservation Planning is responsible for providing the services and functions for monitoring the environment of the OAIS and providing recommendations to ensure that the information stored in the OAIS remains accessible to the Designated User Community over the long term, even if the original computing environment becomes obsolete. Preservation Planning functions include evaluating the contents of the archive and periodically recommending archival information updates to migrate current archive holdings, developing recommendations for archive standards and policies, and monitoring changes in the technology environment and in the Designated Communities' service requirements and Knowledge Base. [...] Preservation Planning also develops detailed Migration plans, software prototypes and test plans to enable implementation of Administration migration goals. (SYSTEMS 2002).

It was obvious to the development team that the missing functionality in RODA would easily be fulfilled by CRiB and its set of components. CRiB, because of its service-oriented nature, would integrate seamlessly with the rest of the components and services developed around Fedora.

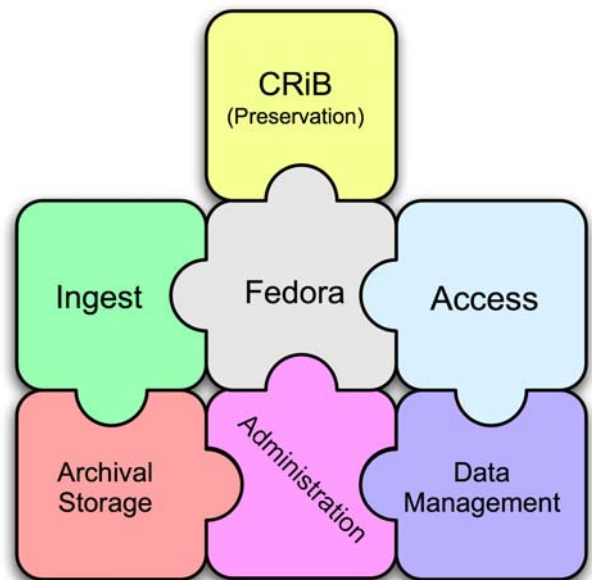


Figure 6: RODA with the new component

In order to fully satisfy the initial requirements of RODA, CRiB would have to be able to handle an additional class of digital representations, i.e. relational databases. Future work will focus on the development of a taxonomy of significant properties for relational databases, the specification of a long-term preservation format/schema for this class of objects, the development of migration services for distinct database products (e.g. Oracle, SQL Server, PostgreSQL, MySQL and others). Some groundwork on this subject has already been initiated and may be consulted at (Ramalho et al. 2007; Henriques et al. 2002).

References

- Barbedo, F. 2006. Especificação de requisitos. Technical Report 41012-005, IAN/TT.
- COMPANY, H.-P., and LIBRARIES, M. Dspace web site. <http://www.dspace.org>. <http://www.dspace.org>.
- Curtis, J.; Koerbin, P.; Raftos, P.; Berriman, D.; and Hunter, J. 2007. Aons - an obsolescence detection and notification service for web archives and digital repositories. *New Review of Hypermedia and Multimedia* 13.
- Darlington, J. 2003. Pronom - a practical online compendium of file formats. *RLG DigiNews* 7.
- Ferreira, M.; Baptista, A. A.; and Ramalho, J. C. 2006. A foundation for automatic digital preservation. *Ariadne*.
- Ferreira, M.; Baptista, A. A.; and Ramalho, J. C. 2007. An intelligent decision support system for digital preservation. *International Journal on Digital Libraries*.
- Group, P. W. G. O. O. C. L. C. . R. L. 2005. Data dictionary for preservation metadata: final report of the premis working group oclc online computer library center & research libraries group. Technical report, Dublin, Ohio, USA.
- Henriques, M.; Libreotto, G.; Ramalho, J.; and Henriques, P. 2002. Bidirectional conversion between xml documents and relational data bases. *International conference on CSCW in design*.
- Lagoze, C.; Payette, S.; Shin, E.; and Wilper, C. 2005. Fedora - an architecture for complex objects and their relationships. *Journal of Digital Libraries*.
- of Congress, T. L. 2002. Página oficial do ead versão de 2002. <http://www.loc.gov/ead/>. <http://www.loc.gov/ead/>.
- of Congress, T. L. 2006. Mets webpage. <http://www.loc.gov/standards/mets>. <http://www.loc.gov/standards/mets>.
- Project, F. The fedora digital object model. <http://www.fedora.info/download/2.0/userdocs/digitalobjects/objectModel.html>. <http://www.fedora.info/download/2.0/userdocs/digitalobjects/objectModel.html>.
- Ramalho, J. C.; Ferreira, M.; Faria, L.; and Castro, R. 2007. Relational database preservation through xml modelling. In *Extreme Markup Languages 2007, Montreal - Canada*.
- SYSTEMS, C. C. F. S. D. 2002. National Aeronautics and Space Administration.

Persistent Identifiers distributed system for Cultural Heritage digital objects

Emanuele Bellini *, Chiara Cirinnà *, Maurizio Lunghi *,

Ernesto Damiani**, Cristiano Fugazza**

*Fondazione Rinascimento Digitale
Via Bufalini, 6, 50100, Florence, ITALY
{bellini,cirinna,lunghi}@rinascimento-digitale.it

**Information Technology Dept.
University of Milan
Via Bramante, 65, I-26013 Crema (CR), ITALY
{damiani,fugazza}@dti.unimi.it

Abstract

In this paper, we present a prototype with a novel resolution architecture for an URN based Persistent Identifiers (PI) system in Italy. We describe a distribute approach for implementing the NBN namespace system and illustrate the solutions adopted for the assignment and resolution of the identifiers with the hierarchical and peer-to-peer request forwarding.

Starting from the core motivations for 'persistent identifiers' for digital objects, we draw up a state of art of PI technologies, standards and initiatives, like other NBN implementations. The prototype is still under development and we present the next steps, in particular we describe the interoperability perspective that already partially foresees the NBN prototype.

Introduction

Persistent identification of Internet resources is an important issue within the life cycle approach to cultural and scientific digital library applications, not only to identify a resource in a trustable and granted way, but also to guarantee continuous access to it. It is well-known that Internet resources have a short average life; their identification and persistent location poses complex challenges affecting both technological and organizational issues, involving access and citation of cultural and scientific resources. The use of URLs can not be considered a reliable approach due to the structural instability of links (ex. domains no longer available) and related resources relocation or updating. The current use of the URL approach increases the risk of losing cultural documents or under-using available cultural collections. The issue is more organisational than technical. There are already some initiatives aiming at stabilising Internet addresses, for example setting up a central registry with a

stable reference/name of a resource with a redirect to the actual URL. But for us this is simply not enough. In the Cultural Heritage (CH) domain it is essential not only to identify a resource but also to guarantee authenticity, credibility and continuous access to it.

In synthesis, a first essential component in realising a 'long-term availability' is the use of Persistent Identifiers (PI) in order to solve the problem of univocal identification and **reliable locator** of Internet resources. But another key component to implement a PI service is the **credibility** and long term sustainability of the **Registration Authority**, the institution that stands security for the maintenance of the PI-URL association register, and granting for the resource authenticity, completeness and the content accessibility. Another element to be taken in consideration creating a name space for any type of resources is the level of **service** and '**granularity**' that the identifiers are requested by the specific user application.

Persistent identifiers solution

A trustworthy solution in the CH is to associate a Persistent Identifier (PI) to a digital resource certifying in some way its content authenticity, provenance, managing rights, and providing an actual locator.

Persistence refers to the permanent lifetime of an identifier. It is not possible to reassign the PI to other resources or to delete it. That is, the PI will be globally unique forever, and may well be used as a resource's reference far beyond the lifetime of the identified resource or the naming authority involved. Persistence is evidently a specific matter in a cultural institution's service or policy. The only guarantee of the usefulness and persistence of identifier systems is the commitment shown by the organizations who assign, manage, and resolve the identifiers.

Each PI system foresees the existence of a Registration Authority (RA). The RA is an independent authority that assigns names and guarantees their uniqueness and persistence. Finally, the service tailored on user needs, a naming resolution service corresponds to every naming authority and carries out the name resolution.

These are the main steps to be performed in order to implement a PI system:

- 1) Selection of resources that need a PI and define the level of granularity requested by the user application.
- 2) Identification of a RA suitable/trustable for the digital content and the specific user application. A business model sustainable must be defined.
- 3) Definition of the level of service for resolution of names, in particular the resource info data presented, the rights and access modalities.
- 4) Execution of resource name creation and assignment of one or more URLs in the system register.
- 5) Execution of a resolution service for couples PI-URL.
- 6) Maintenance of the register that associates PI-URL and guarantees of continuous access to the resources.

The first three steps are prerogative of each cultural institution or user application manager, whereas the steps thereafter can be delegated to other authorities, in order to guarantee better economic and functional sustainability of the service.

State of the Art

At present some technological solutions (e.g. DOI, ARK, Handle system, URN) have been already developed but no general agreement has been reached among the different user communities so far: this scenario shows that it is not viable to impose a unique PI technology. Moreover the granularity, that refers to the level of detail at which persistent identifiers need to be assigned, is widely different in each user application sector.

Among existing standards for PIs, the more relevant seem to be the following: Uniform Resource Names (URN), Life Science Identifiers (LSID), Persistent URL (PURL), Archival Resource Key (ARK), Handle System with its Digital Object Identifier (DOI) implementation, and the Library of Congress Control Number (LCCN).

URN is a key standard issued by the IETF and experts are promoting that as a meta name-space in order to include other identification systems. PURL is simply a redirectable of URLs and it's up to the system-manager implement some policies for authenticity, rights, trustability. LCCN is something similar but with a credible policy for trustability and stability of identifiers. The DOI system, is a business-oriented solution widely adopted by the publishing industry and that provides administrative tools and a DRM System. ARK provides peculiar functionalities that are not featured by the other PI schemata, e.g., the capability of separating the univocal identifier assigned to a resource from the potentially

multiple addresses that may act as a proxy to the final resource. Furthermore, we may also find multiple, proprietary implementations for a given schema: the URN-based schema grounded on NBNs has been registered and adopted by the Nordic Metadata Projects but is being separately implemented by individual systems with no reference implementation enabling coordination of information sources.

The Uniform Resource Name Approach

The purpose of a Uniform Resource Name (URN – RFC1737) is to provide a globally unique, stable, location-independent resource identifier which can be used for identification, for access to resource characteristics or for access to the resource itself. The URN specification is part of the IETF family of specifications encompassed by the Uniform Resource Identifier (URI) framework. This framework also includes URLs, which specify both a protocol and a location in order to give access to resources on the web. IANA (Internet Assigned Numbers Authority) is the Registration Authority (RA) for URN namespaces. URNs are designed to enable heterogeneous namespaces mapping and currently, experts are promoting this standard as a common level of integration/interoperability with other 'traditional' identification systems like ISBN-ISSN-SICI (see RFC2288, RFC3044, RFC3187).

Unlike URLs, URNs are not directly actionable (browsers generally do not know what to do with a URN) because they have no associated global infrastructure that enables resolution (such as the DNS supporting URL). Although several implementations have been made, each proposing its own means for resolution through the use of plug-ins or proxy servers, an infrastructure that enables large scale resolution has not been implemented. Moreover, each URN name-domain is isolated from other systems and, in particular, the resolution service is specific (and different) for each domain.

NBN namespace and on-going projects

The National Bibliographic Number (NBN – RFC3188) is a namespace used by National Libraries and based on the standard URN by the IETF. The NBN namespace, as a Namespace Identifier (NID), has been registered and adopted by the Nordic Metadata Projects on request of the CDNL and CENL.

The RFC 3188 says:

'The NBN is a generic name referring to a group of identifier systems utilized by the national libraries and only by them for identification of deposited publication which lack an identifier, or to descriptive metadata (cataloguing) that describes the resources'.

Each National Library uses its own NBN string independently and separately implemented by individual systems with no coordination and no common formats with other national libraries. In fact, several national libraries have developed their own NBN systems for national and

international research projects; several implementations are currently in use, each with different metadata descriptions or granularity levels. An example is the DIVA project at the Uppsala University Library in Sweden, where documents published in the DIVA-Portal have a unique identifier. In cooperation with the Royal Library of Sweden, they implemented an URN-NBN system. One can access every document registered on the DIVA system from the NBN resolver at the Royal Library, whether it may be located at its originating institution or at the Royal Library archive. A similar example is the EPICUR Project at the Deutsche Nationalbibliothek. The aim of the project is to enhance the existing URN-NBN system in Germany for online theses and other types of resources.

There are some important initiatives at European level like the TEL project that it is in the process of implementing a unique system based on NBN namespace within the European Digital Library (EDL). The adoption of NBN identifiers is needed for implementing the 'National Libraries Resolver Discovery Service' as described in the CENL Task Force on Persistent Identifiers, Report 2007.

The NBN Project in Italy

The project, funded by the Fondazione Rinascimento Digitale (FRD) and developed together with the National Library in Florence (BNCF), the University of Milan (UNIMI), and the University consortium (CILEA), has developed a prototype for a national register of digital cultural resources. The first phase of the project has already been completed and the first results are available; future objectives are defined looking for international cooperation. The NBN project is based on a 'trusted digital repository' installed within another project jointly developed by the FRD and BNCF.

First phase objectives

The main objectives of the project first phase have been the following:

- to create a national stable and certified register of digital objects in use by cultural and educational institutions;
- to allow an easier and wider access to the digital resources produced by Italian cultural institutions, including material digitised or not yet published;
- to encourage the adoption of long term preservation policies and make costs and responsibilities for the service sustainable;
- to test a new technology based on URN but upgraded in its architecture with distribution of responsibility for names management;
- to create some redundant mechanisms both for duplication of name-registers and in some cases also for the digital resources themselves.

Second phase objectives

The main objectives of the project second phase will be the following:

- to extend as much as possible the adoption of the NBN technology and the user network in Italy;
- to reinforce the peer-to-peer resolution service and the robustness of the network for direct access to digital resources;
- to develop a protocol for inter-domains (e.g., NBN Italy and NBN Germany, or NBN Italy and DOI) resolution service with a common format of info-data and a friendly user interface.

Hence in the CH context, it is necessary to implement a service for URN assignment and resolution on the national level (managed by the National Library of Florence - BNCF), based on NBN. The decision to utilise the NBN is due to the fact that it is a namespace for the exclusive use of national libraries (every country has registered a sub-namespace at the Library of Congress: for Italy: NBN:IT ISO 3166); this guarantees the presence of the requisites of stability and permanence necessary for an institution that intends to manage a PI service.

The project has developed a prototype that, independently from the content management systems of the single cultural institutions, realises a national register of persistent identifiers for the digital cultural objects on the Internet, and experiments a service of resolution and access to these resources by inserting several elements of novelty in the system's architecture and functionality with respect to the technological solutions currently proposed or under development.

This solution is conceived especially for those resources that do not have any type of identification (i.e. doctoral theses, digitisation of antique books, etc), but in perspective, it can also be extended to unifying all digital cultural resources under a single code, even those are already identified by codes like ISBN, ISSN, SICR, or DOI. The identified resources will thus be able to reside on the system of the cultural institutions that have the rights to manage their sub-domains, and in the legal deposit system. Therefore the expected impact of the project will be to extend, as thoroughly as possible on the national level, the adoption of this technology of stable addressing of the Internet resources in the sphere of culture and education, valorising the scientific and cultural production of the Italian institutions and improving their impact on the public, including that of works that are either little-known or out of print.

A Distributed Approach

A PI distributed system foresees that the responsibility of generation and resolution can be delegated to other institutions called sub-naming authorities, who manage a portion of the name domain/space.

The Italian prototype implements a new PI architecture: the approach is based on URN/NBN, with additional features and solutions recalling the DNS architecture. The prototype defines a hierarchical distributed system, in order to face the criticality of a centralised system and to reduce the high costs of management for a unique resolution service preserving the authoritative control. The project foresees a distributed authority and responsibility for the creation and resolution of names and redefines the central point role, from a unique name generation/resolution access point to an identifier validator and resolution request router. The central node (BNCF) manages the entire domain NBN:IT, but delegates some second-level agencies to manage sub-domains (e.g., NBN:IT:FRD) both in terms of generation and resolution of names.

In this approach the level of resolution trustability increases if the number of institutions joining the network grows up, because, differently from other approaches, in this pilot there is not a single URL for accessing the NBN resolution service. Every node resolves every NBN item generated inside the sub-namespace IT.

The responsibility distribution joined, in some cases, with duplication of data (names+resources) help increase robustness and performance of the system.

NBN System Architecture – Elements & Functions

The architecture of this NBN system is carried out on two levels, with five elements and four basic functions.

ELEMENTS

- 1) central node (BNCF)
- 2) sub-domains register
- 3) second level agencies (cultural institutions)
- 4) NBN sub-domain register
- 5) NBN central register

MAIN FUNCTIONS

- 1) creating a sub-domain
- 2) generating a name
- 3) updating the NBN registers
- 4) resolving a query for a name

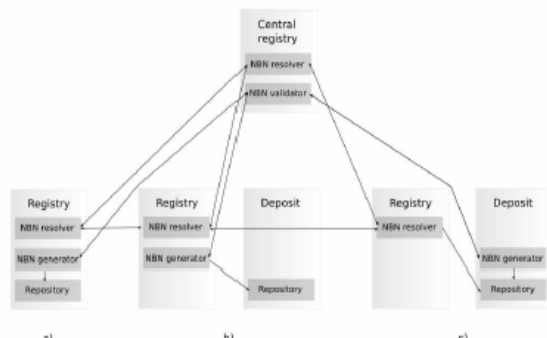


Figure 1 – NBN architecture

Central point

The architecture identifies a central point, located at the National Library of Florence (BNCF) the Registration Authority for the Italian NBN domain, and some second-level institutions. The central node can generate names and sub-domains; it can resolve a user-query directly or redirect it to the appropriate second level agency.

The central node acts in some cases as ‘legal deposit’ archiving also the digital resources.

The system is designed to separate the resolution service from the deposit of the resources.

Sub-domains register

Each second level node is identified by a sub-namespace expressed through the NBN name (for example NBN:IT:BNCR:xxx-xxxx for the National Library of Rome). This registry holds the associations between the sub-namespaces and the base URL of the second level registered institutions. This register is located in the central node for the harvesting function, as well as in the second level nodes for allowing the peer-to-peer resolution process.

Second level agencies

The second level nodes manage their sub-domains, like a DNS, generating names for resources on user demand, keeping a sub-domain register updated with all the associations NBN-URL for their sub-domain names. Most of them offer also a resolver service with a web interface: for names belonging to their sub-domain they are able to solve and provide direct access to the digital resources, for other names they ask the central node to resolve the query or try in peer-to-peer with other second level agencies.

NBN sub-domain register

It is specific for each sub-domain and list the names registered by the second level nodes with all the associations NBN-URL for their sub-domain names.

NBN central register

The central node harvests in OAI-PMH each sub-domain register to check the new names, avoid duplication of names for the same resources, and updated the central register with all the associations NBN-URL for the entire Italian domain. In some cases, it may also have a copy of the digital resource itself, creating a double URL association for that name.

In order to avoid the management costs, the register does not include the descriptive metadata of resources, but only the administrative metadata for managing NBN name’s lifecycle and an external pointer to authoritative metadata belonging to existing institutional repositories.

Creating a sub-domain

The BNCF can generate a sub-domain for any authorised institution providing it with a prefix like NBN:IT:FRD and include this in the sub-domain register that will be also

redistributed to all the other second level nodes. The central node checks periodically the status of the second level agencies and the new names generated.

Generating a name

A name can be generated on user demand, by the central node or by any of the second level agencies, but when a new NBN is generated by institutions, it is not immediately resolvable: an answer is expected from the central point to check uniqueness of name-resource combination. Names are not reusable or changeable. The NBN central register lists all the names within the NBN:IT domain.

Updating the NBN registers

The central point is composed of the central register where there are stored all NBN names generated from any second level institution, a check module of the NBN harvested and a sub-domains register with the URLs of all the second level agencies. We have already seen that the central node is responsible for updating and distributing the sub-domains register. About the names, the central point harvests periodically the NBN records from the second level nodes, then an automated process verifies if the NBNs harvested are correct (see Registries synchronization). Finally the central point sends an answer where are pointed out the NBNs that are not correct or a simple confirmation message if there are no problems, and of course it updates the NBN central register.

Resolving a query for a name

Any second level agency may have installed the resolver service through a simple web page for any name of the NBN:IT domain. If the name requested by the user belongs to the same sub-domain the second level node resolves directly the query, otherwise asks the central node or other agencies in peer-to-peer. The answer is both some info-data and the direct link if access-rights are available. In future, the same web page will be able to resolve also names belonging to other NBNs or to DOI.

Other Functionalities in deep

NBN administrative metadata format

```
<xs:schema>
<xs:element name="nbn-record">
<xs:complexType>
<xs:sequence>
<xs:element name="URI" minOccurs="1"/></xs:element>
<xs:element name="URL" minOccurs="1"
maxOccurs="unbounded"/></xs:element>
<xs:element name="metadataURL" minOccurs="0">
</xs:element>
<xs:element name="MD5" minOccurs="1"/></xs:element>
<xs:element name="creationDate" type="xs:dateTime"
minOccurs="1"/>

```

```
<xs:element name="lastModified" type="xs:dateTime"
minOccurs="1"/>
<xs:element name="status"/></xs:element>
<xs:element name="event" minOccurs="0"/></xs:element>
</xs:sequence>
</xs:complexType>
</xs:element>
</xs:schema>

```

The field creationDate and modifiedDate are used by harvesting engine in order to perform a differential harvesting. In particular the MD5 is a hash field calculated for the physical digital object with the MD5 algorithm. This field is very important for the central point because allows to check if a resource has multiple identifiers. The field status and event are used to track the NBN life cycle as described here below.

NBN life cycle

The project foresees to track each event that may affect the NBN identifier record. There are several “actions” that are managed, like NBN creation or NBN record update. The tracking of update action is important when the resources change their location on the net and consequently change their URL. Another important update action takes place when a new URL is added to for the multi URN-URLs association. The details on the life cycle of NBN identifiers are rendered in Fig 1 as a finite state automa:

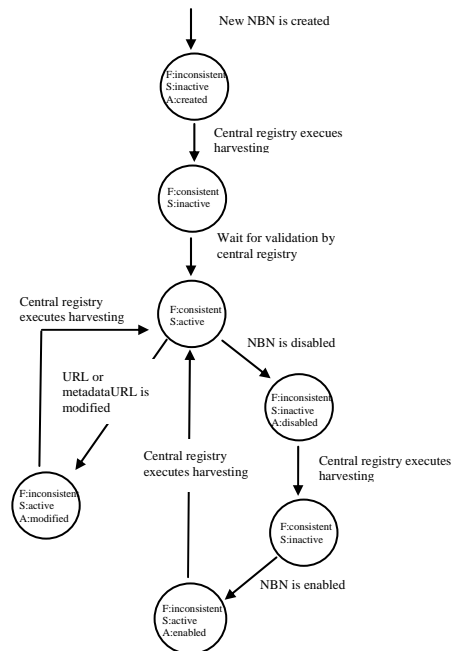


Figure 2 – NBN life-cycle state automa

The distinct states can be singled out by using three variables:

- Flag F, whose value is either 'inconsistent' or 'consistent', that determines whether or not the central register should harvest the record associated with the NBN because it has changed.
- Status S, whose value is either 'active' or 'inactive', that indicates whether registers should resolve the NBN, that is, if the resource associated with the NBN is currently available.
- Action A is an additional variable indicating to the central register, during metadata harvesting, the particular operation that has been carried out (allowed values are 'created', 'enabled', 'disabled', and 'modified').

Registries Synchronization

The architecture foresees the synchronization between central register and second level registries through OAI-PMH protocol. The central node manages the register of sub-namespaces necessary for harvesting of metadata from second level nodes.

This process has 3 steps:

- 1) Harvesting NBN records
- 2) Check NBN records
- 3) Answer to second level register

Harvesting NBN records

The first step is a differential harvesting of NBN records from second level registries. Only the new NBN or NBN records affected by an update will be harvested.

Check NBN records

The second step is the check the data consistency.

Case of alert:

- a) Different NBN and same MD5

An identifier must be assigned to a single resource. If there are other copies of the same resource, the system manages the multi association URN-URLs. The institution that has tried to generate an NBN for a resource that has already an NBN receives a message indicating the right name to be used for that resource.

- b) Identical NBNs and different MD5

- c) Identical NBNs, MD5 and lastModified

These two cases are errors that could happen for many reasons. The prototype sends an alert to the responsible of the last harvested NBN, in order to check and face the problem.

Answer to second level node

The third step is to send the check results to the institution in order to manage the inconsistency. If the problem is the a) case, the institution should disable its NBN identifiers (that will be enabled for another registration). Another email is sent to the owner of the NBN-URL-MD5 first-registered with the new URL of the copy of the same resources included. The owner adds this new URL in the

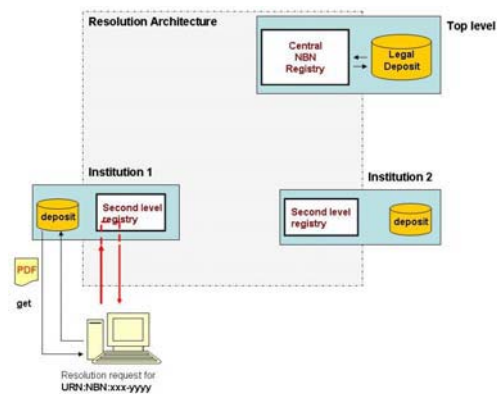
field URL of their NBN registry. The next harvesting and identifier check will enable this multiple resolutions. If the problems are cases b) or c), other activities should be planned like software debug or check for a human error.

NBN Resolution Process

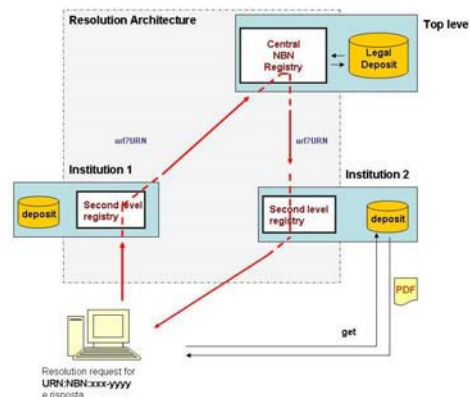
The name resolution request can be submitted by the user to any resolution service of the second level nodes. If the sub-namespace identifies the institution to which the request is submitted, the answers is given directly, otherwise the central registry will be invoked to redirect the resolution request to its appropriate second level node. This architecture increases the robustness of the service and also foresees a peer-to-peer resolution between the second level institutions, in order to maintain the resolution infrastructure operational, even if the central service is not available.

The cases of interaction are the following:

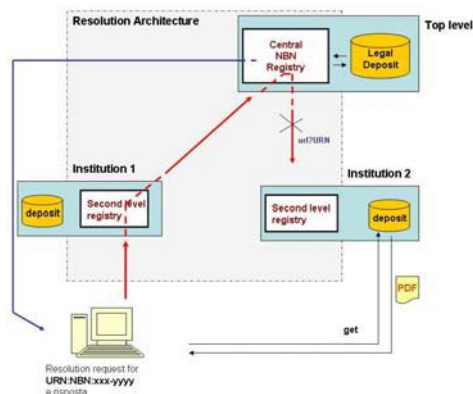
- a) the sub-namespace identifies the institution to which the request is submitted. If the resolution request of this name URN:NBN:IT:FRD:xxx-xxxxx is submitted to FRD resolution service, the answers is given directly.



- b) the sub-namespace does not identify the institution to which the request is submitted. If the resolution request of this URN:NBN:IT:FRD:xxx-xxxxx is submitted to BNCF resolution service, the central registry is invoked to redirect the resolution request to its appropriate second level node.



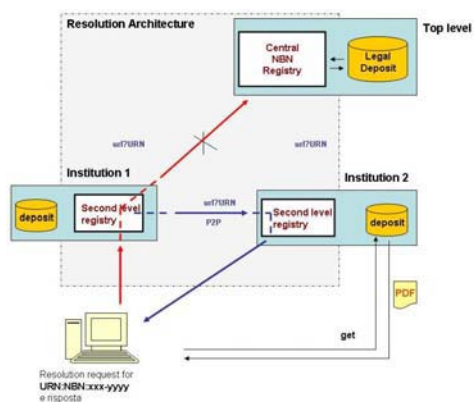
c) the sub-namespace does not identify the institution to which the request is submitted and the appropriate second level node does not work: the central registry answers on behalf of the second level node.



d) the sub-namespace does not identify the institution to which the request is submitted and the central registry does not work: the second level node activates the peer resolution. This is a one of the key features of the entire architecture. This solution can be used as a resolution safe mode or could be selected by a load balancing strategy. In fact a specific load balancing service could decide to forward the resolution request to the peer to peer channel or to the hierarchical resolution process every time.

Peer to peer resolution process

The developed architecture foresees a peer resolution of NBN identifiers. This solution is useful when the central resolution service for any reasons does not work. This feature is necessary because the hierarchical approach has still a single point of failure for a full resolution service. When the central point does not work, every second level institutions are able to resolve their NBN identifiers only without a peer resolution system. The trustability of peer resolution depends on what synchronization strategies are adopted to line up the second level sub-namespace registry with central point sub-namespace registry. Each second level point has a copy of the sub-namespace registry of the central point. The second level resolution service is able to recognize the sub namespace of the NBN string and forward this resolution request to the appropriate second level institution, using the sub-namespace registry. The use of peer-to-peer resolution as a back up service of the hierarchical resolution is a choice. In fact is possible to set the peer resolution as primary and call the central register only if there is no answer from peer or implementing a load balancing service as described above.



Interoperability approach

The central node allows interoperability functionalities with other namesystems, included other NBN systems, as well as the DOI system. The PI systems are thought as autonomous systems. The NBN project has designed the central node as a gateway to forward towards other domain (NBN:DE, DOI, ARK) the resolution request of other NBN namespace identifiers. This approach is a first step for a wider interoperability project among different PI domains for a common resolution service. This function is under development.

Ongoing research activity

The FRD in conjunction with mEDRA (European DOI Registration Authority), CINECA, CILEA, the University of Milan, the central Library of National Research Council (CNR) in Italy are developing a common base resolutions service with DOI, in order to realise a full interoperability with these two identification systems. The approach follows also the CENL recommendations as the 'last resort resolutions' of DOI by NBN. Another important development of the pilot is to establish a common resolver service with other NBN systems in other countries.

Outcome and expected impact

The expected impact is not only a great improvement in the quality of the web coverage of European cultural resources and the reduction of costs and efforts needed to maintain a stable reference of Internet resources, but also a general increasing of credibility and trustability for digital libraries, by promoting the use of digital contents in different user sectors and applications. In particular:

- Persistent identifiers and NBN promotion
- Prototype development and technology evaluation
- Open Source technologies promotion
- Digital preservation development
- Preservation of the minor literature
- Access to resources of difficult or impossible search

References

- J. Hakala. Using national bibliography numbers as uniform resource names, 2001. (RFC3188)
<http://www.ietf.org/rfc/rfc3188.txt>.
- J. Kunze. The ARK Persistent Identifier Scheme. Internet Draft, 2007.
<http://tools.ietf.org/html/draft-kunze-ark-14>.
- C. Lagoze and H. V. de Sompel. The Open Archives Initiative Protocol for Metadata Harvesting, version 2.0. Technical report, Open Archives Initiative, 2002.
<http://www.openarchives.org/OAI/openarchivesprotocol.html>.
- Norman Paskin. *Digital Object Identifiers*. Inf. Serv. Use, 22(2-3):97-112, 2002.
- Kathrin Schroeder. *Persistent Identification for the Permanent Referencing of Digital Resources - The Activities of the EPICUR Project Enhanced Uniform Resource Name URN Management at Die Deutsche Bibliothek*. The Serials Librarian, 49:75-87(13), 5 January 2006.
- Sam X. Sun. Internationalization of the Handle System - A persistent Global Name Service. 1998.
<http://citeseer.ist.psu.edu/sun98internationalization.html>.
- DCC Workshop on Persistent Identifiers
30 June – 1 July 2005 Wolfson Medical Building, University of Glasgow
<http://www.dcc.ac.uk/events/pi-2005/>
- ERANET workshop Persistent Identifiers
Thursday 17th - Friday 18th June 2004-University College Cork, Cork, Ireland
<http://www.erpanet.org/events/2004/cork/index.php>
- Object Management Group. Life Sciences Identifiers final adopted specification, 2004. <http://www.omg.org/cgi-bin/doc?dtc/04-05-01>.
- Dublin Core Metadata Initiative. Dublin Core Metadata Element Set, Version 1.1.
<http://dublincore.org/documents/dces/>.
- H.-W. Hilse, J. Kothe *Implementing Persistent Identifiers: overview of concepts, guidelines and recommendations*, 2006, ix+57 pp. 90-6984-508-3
<http://www.knaw.nl/ecpa/publ/pdf/2732.pdf>
- E. Bellini, C. Cirinnà, M.Lunghi, 2008, *Persistent Identifiers for Cultrual Heritage*, DigitalPreservationEurope Briefing Paper
http://www.digitalpreservationeurope.eu/publications/briefs/persistent_identifiers.pdf
- E. Bellini, M.Lunghi, E. Damiani, C. Fugazza, 2008, *Semantics-aware Resolution of Multi-part Persistent Identifiers*, WCKS 2008 conference.
- CENL Task Force on Persistent Identifiers, Report 2007
http://www.nlib.ee/cenl/docs/CENL_Taskforce_PI_Report_2006.pdf.
- National Library of Australia, PADI (Perserving Access to Digital Information) Persistent Identifiers , 2002
<http://www.nla.gov.au/padi/topics/36.html#article>
- Relationship Between URNs, Handles, and PURLs
Library of Congress, National Digital Library Program
<http://lcweb2.loc.gov/ammem/award/docs/PURL-handle.html>
- Corporation for National Research Initiatives *Handle System (Overview)*
<http://www.handle.net/overviews/overview.html>
- Coyle, Karen Identifiers : *Unique, Persistent, Global*
<http://dx.doi.org/10.1016/j.acalib.2006.04.004>
Journal of Academic Librarianship, Volume 32, Issue 4, (July 2006), p. 347-450
- Bermes, Emmanuelle, International Preservation News, Vol 40 December 2006, pp 23-26 *Persistent Identifiers for Digital Resources: The experience of the National Library of France*
<http://www.ifla.org/VI4/news/ipnn40.pdf>
- Andersson, Stefan; Hansson, Peter; Klosa, Uwe; Muller, Eva; Siira, Erik Using XML for Long-term Preservation : Experiences from the DiVA Project
- David Giarretta, Issue 1, Volume 2 | 2007
The CASPAR Approach to Digital Preservation
The International Journal of Digital Curation
- Sollins, Karen *Architectural Principles of Uniform Resource Name Resolution* (RFC 2276)
<http://www.ietf.org/rfc/rfc2276.txt>
- Masinter, Larry; Sollins, Karen *Functional Requirements for Uniform Resource Names* (RFC 1737)
<http://www.ietf.org/rfc/rfc1737.txt>
- Moats, Ryan 1997URN Syntax (RFC 2141)
<http://www.ietf.org/rfc/rfc2141.txt>

Encouraging Cyberinfrastructure Collaboration for Digital Preservation

Christopher Jordan (TACC), Ardys Kozbial (UCSDL)**, David Minor (SDSC)***, and Robert H. McDonald (SDSC)***

* Texas Advanced Computing Center (TACC)
J.J. Pickle Research Park, 10100 Burnet Road (R8700), Building 196, Austin, TX 78758-4497
ctjordan {at} tacc.utexas.edu

** UCSD Libraries,
University of California, San Diego – 9500 Gilman Drive MC-0505, La Jolla, CA 92037-0505
{akozbial} at ucsd.edu

*** San Diego Supercomputing Center - University of California, San Diego – 9500 Gilman Drive MC-0505, La Jolla, CA 92037-0505
{minor, mcdonald} at sdsc.edu

Abstract

Over the last several decades, U.S. supercomputing centers such as the San Diego Supercomputer Center (SDSC), the National Center for Supercomputer Applications (NCSA), and the Texas Advanced Computer Center (TACC), along with national partnerships such as the National Partnership for Advanced Computational Infrastructure (NPACI) and TeraGrid have developed a rich tradition of support for advanced computing applications and infrastructure. In addition, these centers have developed some of the worlds longest continually operating archives of digital information. These characteristics enable such nationally-funded centers to become natural partners for the library and archive communities as they develop digital preservation infrastructure. Concepts which will be critically important to the development of long-term preservation networks, including cyberinfrastructure and data grids, have grown out of the National Science Foundation and its programs for supercomputer centers. The centers have also served as hosts for long-running development and testing of software tools for data management in distributed environments, including the SRB and iRODS data grid software. These centers are also natural sites for the deployment of necessary physical and virtualized cyberinfrastructure for digital preservation. Several important current and past initiatives, from InterPARES (Duranti) to Chronopolis have involved staff and resources at supercomputing centers working directly with archives and libraries.

Along with these opportunities, there are significant challenges to the integration of the current infrastructure involved in the support of advanced computational science, on the one hand, and services that support the community needs for digital preservation on the other. This paper provides an overview of software development and deployments in the context of supercomputing centers and national partnerships, describing foundational

cyberinfrastructure efforts, which provide physical and logical support for more advanced digital collection and preservation projects in both the sciences and the humanities. The paper then surveys some important recent work at sites in the NSF's national cyberinfrastructure project, the TeraGrid, related to the digital preservation arena. It also examines two projects that the Library of Congress' National Digital Information Infrastructure and Preservation Program Program (NDIIPP) has funded at SDSC to study large-scale, long-term digital archives. These projects provide valuable examples of collaborative digital preservation practice within the context of a shared U.S. cyberinfrastructure.

Finally, we consider the possibilities for further development of digital preservation infrastructure and partnerships within the Teragrid and across international boundaries. The character of digital preservation development outside of the United States is briefly considered and compared, and future directions for international efforts are evaluated.

Introduction

In recent years, there has been an increasing level of interest and effort on both the intellectual and practical aspects of digital preservation in the information technology and open science communities. Commercial, non-profit, and government entities have all produced reports and funded investigative efforts on various aspects of the problems of data management and long-term digital preservation. This paper argues that the efforts of institutions as diverse as libraries, museums, science and engineering funding agencies, and supercomputing centers are properly seen as complementary, although these institutions may not have long histories of collaboration, and have seemingly focused on very different disciplinary activities in the past. Further, we argue that continuing efforts to engage in collaborative relationships across institutional and disciplinary boundaries have already begun to bear fruit,

and should be further encouraged as the theory and practice of digital preservation matures.

Supercomputing Centers as Cyberinfrastructure Laboratories

The U.S. National Science Foundation (NSF), working with public and private research universities across the U.S. over the last several decades, has built a broad portfolio of institutions dedicated to the use of computational resources to enable open science research. Some of the largest current examples of these institutions include the National Center for Supercomputing Applications(NCSA) at the University of Illinois, the San Diego Supercomputer Center(SDSC) at UC San Diego, and the Texas Advanced Computing Center(TACC) at the University of Texas. More recently, NSF has funded multiple phases of a national partnership, currently known as the TeraGrid and including a total of 11 institutions, to further facilitate the use of scientific computing resources by the national community of researchers(Berman). Historically, these centers have organized their mission around the provision of large and expensive supercomputers with capabilities and resource requirements orders of magnitude greater than typical desktop systems. However, due to the complexity of the tasks involved in utilizing supercomputers, and the sophisticated infrastructure required to support computational science on systems at this scale, the centers have become the natural location for a wide variety of advanced research activities relying upon or in support of high-end computational science. These research and support activities include high-speed networking, software development, scientific visualization, and data management and archival services.

The breadth and depth of facilities and activities necessary to support current and future research using computational resources was noted in a seminal NSF blue-ribbon panel report (Atkins), and the combination of the physical and human infrastructure was described, in that report, as *cyberinfrastructure*. The same report recommended that the NSF should explicitly provide support for research and production support activities across the entire spectrum of cyberinfrastructure needs. While the supercomputing centers were already providing this full spectrum of support functions, the Atkins report and the consequent formation of an Office of Cyberinfrastructure within the NSF created a more explicit sense that the function of the centers was much wider than simply providing high-end computational capabilities for research scientists.

NSF DataNet

In response to the Atkins report, and the obvious and growing need for widespread research on the challenge of access and preservation for vast amounts of science data, the in late 2007 NSF initiated its Digital Data Preservation and Access Network Partners program, also known as DataNet. This program will fund up to five partners over the course of at least five years to perform research into all aspects of the digital data lifecycle as well as production preservation and access functions for

the growing number of digital data collections created by the U.S. research community. An important aspect of the DataNet call for proposals is that it is explicitly stated that DataNet partners should work in combination with TeraGrid partners to support the research community. This requirement will further the integration of data-oriented research and production infrastructure with high-end computing infrastructure. Some TeraGrid partners, like SDSC, are already participating in digital preservation projects funded by the U.S. National Archives and Records administration, and the Library of Congress. Because of the overlapping demands, and in many cases institutions, it is likely that DataNet and TeraGrid partners will continue to participate in a broad range of activities related to digital preservation, even those not directly related to the sciences traditionally supported by supercomputing centers.

Cyberinfrastructure for Preservation

In the course of developing infrastructure to support high-end computational science over several decades, supercomputing centers have developed numerous practices, software tools, and even physical infrastructure for handling massive amounts of data, often for long periods of time. While these mechanisms were rarely designed explicitly to support digital preservation requirements, as links are formed between practitioners in the library, archival, and information technology communities, we are finding parallels between the needs of diverse research communities, and technologies developed to support high-performance computing needs are finding new usefulness in digital preservation environments.

Data Grids

Not least among the reasons for the importance of collaboration in the practice of digital preservation is the need for replication and distribution of data. Replication of data provides protection against rare but inevitable failures in the physical and technical systems used to store and access digital data, while distribution to specific locales may be necessary for geographical protection, high-speed access without the latencies associated with long-distance networking, or even to satisfy legal requirements.

Within the field of supercomputing, the value of utilizing networks to distribute computation and data storage has long been recognized, and significant research has been performed into the problems and potential solutions to the problem of managing vast quantities of data across widely distributed resources, potentially on different platforms and usually in different administrative domains.

The Globus Project

Initiated at Argonne National Laboratory, and now a distributed development project involving numerous components developed there and elsewhere, the Globus project has developed several tools for managing data in a grid context(Chervenak). These tools include the

GridFTP mechanism for high-performance data transfer, and several mechanisms, including the Reliable File Transfer service and the Replica Location Service, for managing the movement of large numbers of files across multiple resources. These software packages were developed in the context of serving the needs of scientific computation and the associated data sets, and are widely used in the TeraGrid and other open science projects. However, as will be discussed below, they are also applicable to the needs of digital preservation-oriented projects.

Data Intensive Computing Environments

The Storage Resource Broker(SRB) is the most widely used software tool to be produced by the Data Intensive Computing Environments(DICE) group at the San Diego Supercomputer Center. More recently, this tool has been superseded by the Rule Oriented Data System(iRODS) software(Rajasekar). Both of these packages provide complete suites of data grid functionality suitable for data-intensive computing applications and digital library applications, including virtual namespaces, data replication, and data verification. The DICE group presents a particularly valuable example of the kind of collaboration encouraged herein, as the groups' origins are firmly in the world of scientific computing, as indicated by the name. However, in recent years collaborations with multiple partners on digital preservation projects, including the National Archives and Records Administration and the UCSD Libraries, have led to important innovations in the structure and usage of data grid software. In particular, the iRODS software was developed specifically to aid in servicing the complex policy and management needs of long-term digital repositories, as opposed to the needs of large scientific data collections, which drove the development of the SRB software. The use of SRB and iRODS software in collaborative environments is described in more detail below.

Long-Term Archival Storage

In addition to the software development activities undertaken within supercomputing centers, a little-noticed aspect of these centers is the fact that they already manage some of the longest-lived digital archives in existence today. SDSC, NCSA, and the Pittsburgh Supercomputing center have all been operating consistently since 1985, and in each case these institutions have been operating a digital archive for the duration of their existence. These archives have been through 2 to 3 complete system migrations, and an even larger number of tape media migrations, in each case, yet have preserved data across each of those migrations. All three centers still have access to files with creation dates in the 1980's. This ability to preserve raw data files over the span of decades, utilizing multiple generations of technology, is almost unparalleled outside of commercial settings, if for no other reason than that very few academic or research institutions outside of the computer sciences and engineering have been working with digital data continuously for this long a timespan.

An important caveat to the achievements of supercomputer centers in preserving data for long periods of time is that these centers have generally considered their responsibility for stewardship of the data to be limited to "bit preservation", i.e. the preservation of the raw data files without any institutional engagement with the contents of those files. The research communities responsible for the generation and use of the data files stored in supercomputer center archives are also expected to be responsible for the management of format information, program code for reading and writing the data, translation or recompilation of executables into forms suitable for new generations of computer systems, etc. This points to another important aspect to collaborative relationships with supercomputer centers: these centers can develop expertise in the technical aspects of preservation, which over time may come to include a much larger set of operations than mere bit preservation, but fundamentally they are service organizations for disciplinary researchers, and therefore function most effectively in a technical support role for users or collaborating institutions who are able to provide expertise in the specific aspects of the data being collected and preserved.

This characteristic can function as both strength and weakness; by focusing on the technical aspects of preservation, an institution can effectively support a range of disciplines and functions, which it would be impossible to support in a single, vertically-integrated institution. On the other hand, without the collaborative relationship with external domain experts, the institution is in danger of losing the contextual information that makes the data it preserves meaningful. This lends an imperative color to collaborative relationships between institutions focused on technology and their partners, and creates a need for specific agreements to govern the process of data transfer to and from the partner institution in case the collaborative relationship dissolves.

High-Performance Network Access

One final area of expertise and infrastructure needs to be noted when discussing the value of supercomputer centers as partners for digital preservation – network availability and services. For a large proportion of those institutions and projects engaged in digital preservation activities, the goal is not simply to preserve digital data in an inaccessible archive, but rather to take advantage of the endlessly reproducible nature of digital data to enable wide dissemination of that data to either specific communities or to the public at large. As with other technical aspects of digital preservation, this requires a level of expertise in high-performance networking, as well as a level of access to high-speed networks, which interconnect academic and other institutions.

For the same reasons that supercomputer centers have developed expertise in the software and practices for managing distributed data, those centers have developed expertise in, and possess considerable infrastructure for, serving large quantities of data over high-speed networks. These resources and skills involve not just the networks themselves but the server systems and software

tools required to enable many large-scale transactions to take place utilizing one or more high-speed networks. Supercomputer centers have been instrumental in the development of, and are long time participants in, high-speed research networks such as the National Lambda Rail and Internet2. Outside of the context of these types of networks, bandwidth costs alone could prove prohibitive for institutions interested in the dissemination of large quantities of data to their designated communities.

Libraries in the Digital Age

The data deluge is beginning to have an effect on libraries and archives. As custodians of the scholarly record, libraries and archives are being asked to play an active role in long-term digital preservation in both science and the humanities. A report to the National Science Foundation from the Fall 2006 ARL Workshop on the role of academic libraries in the digital data universe states that “the group found that research and academic libraries need to expand their portfolios to include activities related to storage, preservation and curation of digital scientific and engineering data.” (To Stand, p 42)

One of the major trends in this area is the notion of partnerships, of considering the full set of skills necessary to preserve data for the long term and recognizing that a single group or discipline does not have expertise in all aspects of digital preservation.

Libraries and archives provide expertise in information management, organization and accessibility. Computer scientists and engineers provide expertise in the portfolio of technologies required to support digital preservation. Domain scientists and humanities scholars provide expertise in the content of the data to be preserved. In order to be effective, these groups must work together.

In its partnership, the UCSD Libraries and SDSC have come to realize that collaborative relationships across institutional and sector boundaries “have the potential to spread the burden of digital preservation, create the economies of scale needed to support it and mitigate the risk of data loss.” (Educause, p 10)

TACC and the Texas Digital Library

The Texas Digital Library is an institution founded by the University of Texas at Austin and Texas A& M University, with contributions and participation from several major institutions within the state of Texas. The short-term goal of the Texas Digital Library is to facilitate the creation of Institutional Repositories(Lynch) for the participating institutions, by providing interface and digital library services for those institutions in a framework of cooperative sharing of digital data.

A novel aspect of the TDL efforts is that the consortium is reliant upon the collective participation of its members, with no external funding to provide basic infrastructure services. For this reason, TDL is developing a partnership with TACC to provide storage services, which could be provided for one or more of the participating repositories based on a flexible framework of collective resource sharing. In this partnership model, the supercomputing center provides expertise in the management of archival storage and networking services, while the digital library provides the human and technical interface to the university community or communities, expertise in the ingestion of IR materials, and coordination of the network of participating institutions. Both institutions are able to leverage their collective expertise to provide a service that would otherwise be unavailable due to resource constraints. The demonstration of the importance of this kind of IR service, over time, is expected to lead to steadily increasing levels of institutional commitment, and eventual integration of the concept of the institutional repository into the mainstream understanding of the academic environment within the participating institutions.

The example of TDL and TACC indicates how even in the absence of significant content-specific expertise, supercomputing centers can make significant contributions to the achievement of digital preservation objectives. Simply by providing the basic archival storage infrastructure which is required, supercomputing centers can help institutions to achieve objectives even in disciplines like the humanities, and for services like Institutional Repositories, which would not generally be considered activities engaged in by a supercomputer center. The ability of supercomputer centers to provide reliable storage systems over time spans of decades or more is a significant capability, which when further enhanced by the expertise of digital librarians and domain experts in file formats and contents, can provide a stable foundation for achieving practical digital preservation.

NDIIPP-Funded Projects at SDSC

The Library of Congress’ National Digital Information Infrastructure and Preservation Program (NDIIPP) has funded two projects at SDSC to study large-scale, long-term digital archives: “Data Center for Library of Congress Digital Holdings, a Pilot Project” and “The Chronopolis Digital Preservation Archive and Demonstration.”

Data Center for Library of Congress Digital Holdings, a Pilot Project

This project ran for 18 months, beginning in the summer of 2006. It was described by its PIs as a “trust-building exercise.” Its main goal was to demonstrate how a third part repository (SDSC) could ingest, manage and replicate active digital collections from the Library of Congress. The project worked with two collections: the

complete images of the Prokudin-Gorskii photograph collection from the Prints and Photographs Division, and the complete webcrawl collection from the 2004 Congressional elections. The photograph collection was small in size (about 600GB) but had a complex and unique file structure with parallel usage demands by the Library's staff. The webcrawl collection was large in size (about 6TB) but very uniform in file structure and file types.

First Task: Data Transfer

The first work that SDSC performed for the LC was configuration of a high-speed network connection. The Library had a pre-existing Internet2 connection but was not using it extensively for data transfer, preferring instead to physically transfer data on hard drives. SDSC and LC staff spent significant time configuring connections, including account and security issues, firewall modifications and network tuning, to achieve acceptable throughput. When complete, the team could transfer at a constant rate of 200mb/s, or about 2TB of data per day. While not ideal, this was deemed acceptable for the project.

Another significant part of the data transfer configuration was the use of a transfer tool. Because of their previous grid computing experience, SDSC staff had strong recommendations to use GridFTP. This tool provides the ability to finely tune transfers as well as run them in parallel. It also allows for restarting of interrupted transfers, a key need in this environment. LC staff were less than thrilled with GridFTP as a practical tool in their environment, believing it to be too complex and hard to manage for their needs.

Second Task: Data Replication

The data was managed at SDSC using the Storage Resource Broker. This allowed for multiple active replications of the data to be stored on different storage systems. SDSC stored five copies of the data: two on separate disk systems, two on separate tape archives, and one on a Copan MAID system. Underlying the SRB replication management were two different archiving systems: HPSS and SAM-QFS. This was done as a demonstration of storage diversity and was transparent to the LC staff accessing the data.

SDSC staff also used the SRB to create detailed monitoring and logging scripts to track the data as it moved through the systems and to maintain a high level of reliability.

Third Task: Parallel Webcrawl Indexing

SDSC and LC staff worked together to modify the source code of the Wayback machine software, enabling it to run in parallel on a SDSC cluster. This was done in close conjunction with the Wayback software authors at the Internet Archive, and the changes developed have since been incorporated into the main source tree.

The Chronopolis Digital Preservation Archive and Demonstration

The Chronopolis project began in January 2008, after several years of planning (Moore, 2005). At its core is a nationally-federated data grid housed at SDSC, the National Center for Atmospheric Research in Boulder Colorado (NCAR) and the University of Maryland's Institute for Advanced Computing Studies (UMIACS). Each of these sites is providing 50TB of storage connected via high speed networking. Data for the project will be replicated identically at each of the sites and managed by SRB.

Data Providers

The project is relying on data from the California Digital Library, Inter-University Consortium of Political and Social Research, the Scripts Institute of Oceanography, and the North Carolina State University Libraries. These organizations are providing a wide range of data types, sizes and organizational systems, all of which will be replicated exactly within the Chronopolis framework.

Challenges

This project will focus on several challenges, not the least of which is simply creating such a large data network with active replication. This involves configuring diverse storage systems at the provider sites, high-speed networking for the data transfer, and accurate monitoring of the entire system. Data from CDL and NC State is organized in a new preservation format named "BagIt," and the project will be looking at ways to maximize transfer methodologies for this emerging preservation standard. Data from ICPSR and SIO are stored within SRB and this will form the core of their transfer methods.

The project will also be working with new technologies to monitor nationally-federated collections, including UMIAC's Audit Control Environment (ACE), which provides administrators and data owners detailed views of the status of their data within the system.

Finally, the project is working with metadata librarians from UCSD Libraries and other institutions to create PREMISE definitions for the data and storage systems. This metadata will be created not just to represent the specific data in the system presently, but also to create pathways to other data grids that will come online in the near future and contain similar kinds of collections.

Long-term Goals

This NDIIPP-funded project is an example instantiation of a larger enterprise that SDSC and the other providers view as critically important for future work in digital preservation. The motivating premise is that nationally federated data grids hold an important key to safeguarding and making available digital assets long into the future.

Outlook for Future Efforts

It is clear that the increasing needs of libraries and traditional archives for the preservation and management of very large amounts of digital data, and the natural advantages of collaborative efforts in this context, will continue to drive digital preservation practitioners to search for partners both in and out of the science and engineering disciplines. In this paper, we have described, and demonstrated, the value of cross-disciplinary partnerships in leveraging the expertise and infrastructure of diverse institutions to meet the complex challenges of digital preservation in the 21st century. These efforts, however, are only the beginnings of the necessary efforts to address the size and complexity of digital data as it will be generated and used. Where datasets are 2 terabytes today, they will be 200 terabytes tomorrow, and 2 petabytes the day after that. Where datasets are large image collections today, they will be large image collections with important relationships to textual documents, geospatial data, and moving image data tomorrow. The range of expertise required to ingest, to curate, and to preserve these collections for current and future generations will continue to expand. Information scientists, archivists, computer scientists, and engineers will all have important roles to play in performing these tasks, and it is arguable that no one institution will have the resources to accumulate all the infrastructure and expertise necessary.

We expect that, with the introduction of the NSF's DataNet Partners, the collaborative model will become more common, and that particularly as links are formed to broader cyberinfrastructure partnerships like the TeraGrid, larger and larger collaborative efforts will become the norm. An important aspect for future investigation will be the optimal size and organization for these partnerships, and how multi-tiered institutional mechanisms for the management, preservation, and dissemination of digital data can operate most effectively within regional and national contexts.

Another critical aspect for these partnerships will be the question of where long-term support for the practice of digital preservation will come from. There is a critical need, particularly within the United States, for additional support both from the discipline- or project-specific side and from the infrastructure side. The DataNet proposal explicitly includes as a condition that after ten years, partners must be sustainable independent of NSF funding. Without the development of institutional commitments on the scale and timeframe of those assumed for university libraries, or significant endowments for institutions engaging in digital preservation, it is unclear how the ongoing needs of digital data will be served over the long term.

International directions

In addition to the partnerships across disciplines described here, an area of relatively little investigation up to this point is the potential and outcome of international cooperation, between institutions with similar or diverse specializations. As collaborative efforts at preservation become more the norm than the exception, it is

inevitable, and laudable, that these collaborative efforts begin to take on an international character. There will, however, be significant challenges for these efforts, as funding models, language, and simple geography have heretofore encouraged more localized foci for institutional efforts, leading to divergent practices, standards, and technologies. In addition, U.S. funding dedicated to digital preservation has traditionally lagged behind that available in the European and British contexts in particular, so levels of sophistication and maturity in the efforts being undertaken within various nations will be a challenge. As with all aspects of science and technology research, digital preservation is properly seen as an international effort, with a global audience and a global body of practitioners. It is hoped that the types of efforts described in this paper will be continued at multiple scales to support the ever-expanding needs of digital preservation.

References

- American Council of Learned Societies Commission on Cyberinfrastructure for the Humanities and Social Sciences. *Our Cultural Commonwealth*. (2006)
- Atkins, D., et al. "Report of the National Science Foundation Blue-Ribbon Panel on Cyberinfrastructure" (2003). <<http://www.nsf.gov/od/oci/reports/toc.jsp>>
- Berman, F. "From TeraGrid to Knowledge Grid." In: *Communications of the ACM*, p. 27 (2001).
- Berman, F., A. Kozbial, R. H. McDonald, and B. E. C. Schottlaender. "The Need to Formalize Trust Relationships in Digital Repositories" In: *Educause Review*, p 10. (May/June 2008)
- Chervenak, A., et al. "The Data Grid: Towards an architecture for the distributed management and analysis of large scientific datasets." In: *Journal of Network and Computer Applications* 23, p. 187 (2000).
- Duranti, L., et al. *The Long-term Preservation of Authentic Electronic Records: Findings of the InterPares Project*. Available online <<http://www.interpares.org/book/index.htm>>.
- Lynch, Clifford. "Institutional Repositories: Essential Infrastructure for Scholarship in the Digital Age". In: ARL Bimonthly Report 226, February 2003. Available online <<http://hdl.handle.net/2108/261>>.
- Moore, R.L., J. D'Aoust, R.H. McDonald, and D. Minor. "Disk and Tape Storage Cost Models." In: *Proceedings of the IS&T Archiving Conference*, p. 29 (2007).
- Moore, R.W., F. Berman, D. Middleton, B. Schottlaender, J. JaJa, and A. Rajasekar. Chronopolis. "Federated Digital Preservation Across Time and Space."

In: *Proceedings of the Local to Global Data Interoperability - Challenges and Technologies, Sardinia, Italy*, p. 171-76, (2005).

Rajasekar, A., et al. "A Prototype Rule-Based Distributed Data Management System." In: *Proceedings of HPDC Workshop on Distributed Data Management, Paris, France*. (2006)

To Stand the Test of Time. A Report to the National Science Foundation from the ARL Workshop on New Collaborative Relationships: The Role of Academic Libraries in the Digital Data Universe, p 42. (2006)

Establishing a Community-based Approach to Electronic Journal Archiving: the UK LOCKSS Pilot Programme

Adam Rusbridge*, Seamus Ross**

* Digital Curation Centre
Humanities Advanced Technology
and Information Institute (HATII)
University of Glasgow
a.rusbridge@hatii.arts.gla.ac.uk

** Digital Curation Centre
Humanities Advanced Technology
and Information Institute (HATII)
University of Glasgow
s.ross@hatii.arts.gla.ac.uk

Abstract

Lots of Copies Keep Stuff Safe (LOCKSS¹) represents a sophisticated combination of technical and business-aware elements that can be deployed to ensure the long-term accessibility to electronic journal content even if the publisher ceases to exist, a subscription is terminated, or the already acquired content becomes damaged. Given the potential benefits of LOCKSS to the UK community, and in consideration of the implications of the NESLi2 licences, the Joint Information Systems Committee² and the Consortium of University Research Libraries³ (JISC/CURL) co-funded a UK LOCKSS Pilot Programme to explore issues associated with the practical implementation of LOCKSS in UK Higher Education institutions. The pilot launched in March 2006 and concluded in July 2008. Following on from our experiences throughout the UK LOCKSS Pilot Programme, this paper discusses the organizational attributes of the LOCKSS approach that we expect to further develop in the UK, describes the types of journal content that the current generation of LOCKSS seems best suited to handle and as a result how LOCKSS may fit into the broader journal archiving environment, and it describes the steps we are taking to ensure both the LOCKSS software and Technical Support Service grow effectively to support library use and information management.

The UK LOCKSS Pilot Programme

Lots of Copies Keep Stuff Safe (LOCKSS) represents a sophisticated combination of technical and business-aware elements that can be deployed to ensure the long-term accessibility to electronic journal content even if the publisher ceases to exist, a subscription is terminated, or the already acquired content becomes damaged. The LOCKSS approach provides a critical component in the journal distribution infrastructure, allowing libraries to take custody of the assets for which they have paid, while enabling them to conform to the licensing arrangements they have agreed with publishers they will

adhere to, and sharing the technological infrastructure among the wider UK and global library community. The LOCKSS approach makes certain that libraries are responsible not only for short term access, but also involved at many stages in the emergence of this journal archiving service. The LOCKSS system can help to improve confidence in electronic journals, and could help libraries justify to their academic colleagues a move from mixed print and electronic to all electronic in some cases; eventually providing savings far greater than the cost of participation in the initiative.

The National e-Journals Initiative⁴ (NESLi2) Model License developed by JISC for e-journal subscription agreements includes archiving clauses to provide libraries with some reassurances that they will receive continued access to the content for which they have paid. Practical implementations of the archiving clauses, which involve a collaborative agreement between libraries and publishers, are not yet fully in place. LOCKSS provides participating libraries and publishers with a distributed technical architecture to make certain purchased content remains accessible without a necessary dependency on the publisher's market presence. The LOCKSS model really shines in a collaborative context as its implementation within the UK academic library environment has demonstrated. Given the potential benefits of LOCKSS to the UK community, and in consideration of the implications of the NESLi2 licences, the Joint Information Systems Committee and the Consortium of University Research Libraries (JISC/CURL) co-funded the UK LOCKSS Pilot Programme to explore issues associated with the practical implementation of LOCKSS in UK Higher Education institutions. Running between March 2006 and July 2008, the UK LOCKSS Pilot Programme is described in Hockx-Yu (2006) and Rusbridge and Ross (2007).

Briefly, LOCKSS is a collaborative, library-centric approach to electronic journal archiving. Each institution locally runs a LOCKSS box and collects content according to their individual collection development

1 LOCKSS Website, <http://www.lockss.org>

2 JISC Website, <http://www.jisc.ac.uk>

3 Now known as Research Libraries UK (RLUK), <http://www.rluk.ac.uk/>

4 NESLi2 Website, <http://www.nesli2.ac.uk/>

policies. The two year JISC/CURL funded UK LOCKSS Pilot Programme was intended to investigate the practical issues associated with running LOCKSS in the UK and building an effective Alliance of UK institutional partners, to explore issues associated with making available through LOCKSS a wide corpus of journal content which covers the needs of the UK HE library community, and to develop the infrastructure needed to support institutions participating in the LOCKSS approach. At the beginning of the UK LOCKSS Pilot Programme, we determined that in order to run an effective support service we needed to establish a number of distinct components. Our expectation was that at the end of the pilot, with these components in place, an assessment would be made on the desirability of future use of LOCKSS versus available alternatives. The community would share future running costs and technical support (if still necessary) would be built in to an organisation such as the DCC. We recognised that our ongoing support requirements would need to reflect the needs of the UK community. We aimed to set up an infrastructure that would allow us to easily facilitate dialogue between the support service and individual participating institutions. From the outset, the project has built upon the infrastructure of the DCC for technical and training elements. The programme budget included equipment costs for the bulk purchase of LOCKSS boxes, for two years provision of technical support for librarians and technical software and plugin development.

Initially, twenty-four institutions joined the UK LOCKSS Pilot Programme. A further six institutions joined as associate members in July 2006. Throughout the pilot, we hosted a series of workshops bringing together librarians, JISC, and project staff from the UK LOCKSS Pilot Programme and the US LOCKSS Alliance. These workshops allowed different stakeholders to communicate progress, to allow each participant to understand where difficulties were being encountered and improvements could be made, and to ensure that we could achieve consensus on the overall strategy as we adapted to changing requirements and alongside emerging approaches.

Content Complete Ltd⁵, JISC's negotiation agent for NESLi2 content, undertook content negotiation to bring in to the LOCKSS system electronic journal content from publishers participating in the NESLi2 initiative. Our objective regarding content was 'to build a substantial collection of e-journals to which the participating institutions have archival rights'. Establishing a procedure to make available open access material in the LOCKSS network was of interest and a sub-project entitled OpenLOCKSS⁶ was initiated at Glasgow University to negotiate and make available open-access material.

From August 2008, the UK LOCKSS Alliance will transition from a JISC/CURL funded pilot programme to

a full-fledged national service and as part of this transition it will move from the Humanities Advanced Technology Institute (HATII)⁷ at the University of Glasgow. The UK LOCKSS Alliance will build on the experiences of the UK LOCKSS Pilot Programme and be hosted by EDINA⁸, the UK data centre at the University of Edinburgh, in conjunction with the Digital Curation Centre. All UK Higher and Further Education (HE/FE) institutions are welcome to join the UK LOCKSS Alliance. The UK community will share the costs of running the service and libraries that wish to participate can do so under a JISC Collections banded fee basis.

This paper discusses the organisational attributes of the LOCKSS approach that we expect to further develop in the UK, describes the types of journal content that the current generation of LOCKSS seems best suited to handle and as a result how LOCKSS may fit into the broader journal archiving environment, and it describes the steps we are taking to ensure both the LOCKSS software and Technical Support Service grow effectively to support library use and information management.

Alongside our own internal evaluation throughout the Pilot Programme that will be released this autumn, two recent externally led reports have considered the suitability of LOCKSS in the context of the UK higher education environment. Morrow, *et al* (2008) describes a number of scenarios which suppose that a given publisher is no longer in a position to provide access to electronic content and considers the resultant access as provided through a variety of different journal archiving approaches. Dalton and Conyers (2008) reports on a formal assessment of the UK LOCKSS Pilot Programme, considering the overall success and impact of the Pilot against its original objectives and producing a list of recommendations for the ongoing development and improvement of the UK LOCKSS Community. These reports reach the conclusion that the UK LOCKSS Pilot Programme has demonstrated a way in which an effective LOCKSS Alliance can be established and run, and provides a model for other national, regional, or trans-institutional consortia groups. We are delighted that these independent reviews reached these conclusions, and have decided that it will be valuable to consider in more detail the organisational attributes that our experience of running the UK LOCKSS Pilot Programme leads us to believe will produce a successful and sustainable journal archiving infrastructure.

Developing an Infrastructure for UK Journal Archiving

The risks that threaten long-term access to journal content are numerous and have been well elaborated in a many papers; for example, Rosenthal *et al* (2005) describes the threats a digital preservation system should address. Morrow *et al* (2008) sketches a suite of scenarios that suppose that a particular publisher ceases

⁵ <http://www.contentcomplete.com/>

⁶ <http://www.lib.gla.ac.uk/Research/openlockss/>

⁷ <http://www.hatii.arts.gla.ac.uk>

⁸ <http://www.edina.ac.uk>

to be in a position to provide access to electronic content. In this section, we describe some of the attributes we believe will result in a strong and stable foundation for the ongoing UK LOCKSS Alliance.

Librarians are looking for long-term solutions to the issues that arise from digital distribution and the 'acquisition' of access to digital objects and they seek to invest in stable, long-term systems and infrastructures. To reassure libraries that the LOCKSS initiative is worthy of continued investment it will be necessary to demonstrate sustainability, not just in financial terms,⁹ but also in terms of organisation (for example, by addressing risks that arise from staff turnover or from transition of the organisational responsibility from one lead body to another). Libraries will gradually acquire trust in archiving initiatives as they mature and demonstrate their ability to respond to challenging scenarios. It is though worth examining the contributions the UK community can make in supporting the take up and extension of the initiative.

We believe a core strength of the LOCKSS approach is the collaboration that it fosters between librarians, community bodies such as JISC and CURL, and journal archiving initiatives. The UK LOCKSS Alliance has made first steps towards achieving a financially sustainable approach for the UK by moving away from a grant-funded pilot programme to one where the costs will be met by the stakeholders—that is by those directly benefiting from participation in the initiative. We believe that a crucial next step in guaranteeing a stable, ongoing environment is to ensure that librarians are not just well informed of the architecture, and system operation and content negotiation activities, but actively contribute to the direction and development of the initiative. To achieve this, we hope to build a community organised around core principles to counter some of the more predictable risks that may arise as collaborating organisations act to run a long-term digital preservation initiative.

Share responsibility and governance

Shared governance will foster the development of a UK LOCKSS Alliance that reflects a broad consensus of the UK library and curation communities, and that adapts to meet emerging cultural and technological demands. During the UK LOCKSS Pilot Programme, we found the input of librarians enormously useful and as such we are structuring the UK LOCKSS community so librarians are actively engaged in the initiative as opposed to passive users of its services. LOCKSS is a system embedded within the library organisation and as such should reflect the needs of both library staff working with LOCKSS and patrons using the archived content. By developing a system that closely matches the needs and expectations of the user community, librarians are more likely to

continue to use and value the approach. By basing the model of software development on collaborative and distributed open source principles, system development is not reliant on a single team of individuals or a single organisation's finances, a factor that contributes to mitigating such risks as those that can arise through staff loss or organisational failure. Encouraging active discussion and collaboration is essential if the roles and responsibilities of each of the stakeholders are to be effectively examined, discussed, clarified and agreed. Throughout these discussions there will continue to be intense interest in how LOCKSS compares with other journal archiving initiatives as they emerge and develop. Discussion of the strengths and weakness of other journal archiving systems will provide LOCKSS with an indication of areas for possible development and risk.

Develop UK infrastructure

The UK community should avoid becoming overly reliant on resources outside UK control. In this case the UK curation and library communities need to acquire local knowledge, skills and physical infrastructure for journal archiving. The UK LOCKSS Pilot Programme took steps towards this, providing participants with a working knowledge of the technical and licensing issues associated with collection and preservation of electronic journal content. The support service has developed capacity to contribute to LOCKSS system development and developed a strong working relationship with the LOCKSS Development Team in Stanford. Utilising the expertise of librarians for open access negotiation proved effective within the OpenLOCKSS project and we would like to build on this experience. As we describe in a later section, it is likely that LOCKSS will only cater for a proportion of electronic journal content. Complementary to LOCKSS, there will be alternative approaches focusing on specific types of content and each with their own benefits. Where possible, it seems sensible that the UK community should participate as a collective in these as this will foster a UK-wide sense of shared responsibility for infrastructure and assets.

Develop local collections

Journal archiving initiatives must be monitored if there is to be any certainty that the content being negotiated and preserved is relevant to and in accordance with collection policy contexts of the individual participating institutions. One of the strengths of the LOCKSS approach is that it gives a participating library the ability to determine the content that that institution wishes to preserve. We are exploring mechanisms, described later in this paper, which will provide librarians with appropriate opportunities to identify the broadest range of content they would wish to see targeted for inclusion in UK LOCKSS. We aim to push forward with developing an infrastructure that will provide libraries with the kinds of collections their users need. Working closely with our participating libraries to better understand their collection development needs, plans and trajectories is essential if we are to achieve this goal.

⁹ In 2007, incoming fees from Alliance members covered the costs incurred by the Stanford based LOCKSS team. The Stanford team expects this target will continue to be met in 2008.

The UK LOCKSS Pilot Programme has made first steps towards establishing an environment in accordance with these requirements. Journal archiving will necessitate long-term organisation and management and it seems unwise to rely upon a model where too much responsibility and workload lies with isolated organisations and individuals. Likewise, it would be inappropriate to require individual librarians to undertake onerous activities and procedures. As we move forward, we need to understand which aspects of the LOCKSS approach librarians have found to be successful and where improvements could be made. It is possible that the LOCKSS approach may not be suitable for all institutions. For example, some libraries found that they did not have sufficient resources to manage a LOCKSS box alongside their existing services. Some librarians have indicated they saw inefficiencies associated with the maintenance of multiple archives and that reorganisation of the infrastructure and management to utilise data centres acting on behalf of the HE community might, in their view, lower resource constraints without compromising the benefits of shared responsibility and a semi-distributed architecture. One example of this model is CLOCKSS¹⁰, the sibling initiative of LOCKSS that has established a dark archive of content on behalf of the global community and successfully demonstrated an alternative organisational approach. In theory, there would be nothing to prevent the UK community from establishing an archive similar to CLOCKSS in organisation and structure with a focus on the material currently in LOCKSS. While we might not necessarily promote a move of this kind, and noting that licensing and access issues would certainly arise as a result of such a move, as the UK digital curation community is still at an early stage in terms of infrastructural and system development it is important to (re)assess the variety of options available.

Building collection of journals

The UK LOCKSS Alliance will build on the infrastructure put in place during the UK LOCKSS Pilot Programme and we intend to implement the recommendations from the recent JISC-initiated evaluation report (Dalton, P. and Conyers, A. 2008). Our experiences over the past two years have given us some ideas as to what content we should expect from a LOCKSS network.

Many librarians note that they were motivated to participate in this journal archiving initiative because it allowed them to provide their academics with assurances that a move to an electronic only environment is a safe and stable strategy. As large commercial publishers provide significant quantities of a library's core reading list material, and librarians reflect on the significant and growing proportions of their budgets being directed towards such publishers, librarians are keen to ensure

that the content provided by these publishers is archived in a variety of journal preservation initiatives.

Since their inception, CLOCKSS and Portico have both been notably more successful than LOCKSS in engaging large commercial publishers (for example, both Elsevier and Nature and participating in those two initiatives, but not LOCKSS). At least in the UK, the emergence of CLOCKSS and Portico has somewhat changed the role of LOCKSS. At the beginning of the pilot we anticipated that content from all publishers would be available through LOCKSS, however now we are starting to see LOCKSS as a component in a larger, complementary set of initiatives. Participating in LOCKSS alongside CLOCKSS and/or Portico appears to provide libraries with a balanced approach that enables them to achieve more comprehensive coverage.

With the emergence of CLOCKSS and Portico, it is worth considering the specific role that LOCKSS can play in the journal archiving environment. LOCKSS is particularly suitable for the broad range of journal content material that may be exposed to a relatively high risk but that fall outside the remit of CLOCKSS and Portico. For example, within the Pilot Programme we have been considering the relative risk to which journal content from small, medium and large publishers may be exposed. We will need to establish mechanisms to identify content that is not just of significant scholarly, cultural or resource value, but that is also potentially fragile. At the same time mechanisms are required that facilitate the matching of content corresponding to these latter criteria with the collection development priorities of individual participating libraries. Balancing the content identified and secured through negotiation across the different collection building policy objectives of the participating libraries poses a challenge. The central role of librarians in the development of LOCKSS negotiated and secured content will continue to expand, thus tapping their wealth of experience in making decisions on content acquisition. We would be keen to establish a mailing list, working group, or portal through which at-risk titles can be nominated, and their significance to our research and teaching communities discussed as a key step in reaching consensus on a title's relevance. Subject specialists are well placed to identify titles within particular domains. It may be possible to build these processes into existing library consortia groups, either regional groups such as the North East and Yorkshire Academic Libraries Purchasing Consortium (NEYAL), or national working groups such as the joint Research Library UK/Society of College, National and University Libraries (RLUK/SCONUL) Task Force on Scholarly Information, JISC Journals Working Group, JISC Libraries Advisory Working Group and the JISC Scholarly Communications Group.

Following the recommendation of the Morrow *et al* report, JISC Collections intend to revise the NESLi2 and NESLi2-SMP licenses to require participation by publishers in at least one journal archiving initiative. Embedding archiving requirements within model licences will be instrumental in gaining a higher

¹⁰ CLOCKSS, for Controlled LOCKSS, is a not-for-profit, community-governed dark archive of web-published content. More information is available at <http://www.clockss.org/>.

proportion of publisher participation in preservation initiatives. There are many publishers and titles that are not covered by NESLi2 licenses and the processes of negotiating with these will need to be agreed and coordinated. The OpenLOCKSS project was initiated at Glasgow University Library to negotiate and make available open-access material. The OpenLOCKSS initiative has demonstrated a model that can be used for Open Access titles, however the process has shown that a certain degree of perseverance is required when negotiating with publishers.¹¹ Project staff working on OpenLOCKSS was required to explain the LOCKSS approach and system to publishers who had not previously encountered it, to resolve publisher concerns about the licensing and access arrangements, and to track progress to ensure that overworked publishers were able to complete the required technical work.

As Dalton and Conyers note: *“It was apparent also in the interview with InformaWorld that lack of apparent demand was a major factor in delays in implementation; if there was seen to be a potential demand they would use this as an opportunity to market their membership of LOCKSS to the library community.”*¹² In light of this, we must consider mechanisms so that library demand for participation in preservation services is appropriately conveyed to publishers. Libraries are more likely to demand a publisher is involved in an archiving initiative if the librarians are confident that the proposed initiative is sustainable, viable, and appropriate. In short, librarians must have confidence in the archiving initiative they are supporting. Libraries themselves may wish to consider whether they can establish a policy whereby they require core collections to be archived in at least one of a shortlist of archiving initiatives, of which LOCKSS may be just one option. As highlighted in the evaluation report, attention should be given to ensure librarians can easily identify the content available within the LOCKSS network. EDINA has recently announced that they will be involved in the development of an electronic journal preservation registry service, acting as a single resource that lists each initiative in which a title is archived. Development of this service will be monitored with interest, as it will ease the process by which librarians can identify which titles are not yet available through particular journal archiving initiatives.

While we have found that archiving solutions involving distributed and shared community responsibility have strengths lacking in single institution based solutions, the effort, such as computer system and storage maintenance tasks, that is required from the partners needs to be contained to the minimal necessary. Throughout the pilot we have identified improvements to the user interface that would alleviate some of the required administration effort, and the LOCKSS team in Stanford are currently incorporating these improvements into future releases. Other improvements in the works include such simple changes as bringing clarity to the user interface terminology.

¹¹ <http://www.lib.gla.ac.uk/Research/openlockss/>

¹² See Dalton, P. and Conyers, A. (2008); page 19.

LOCKSS Technical Support Service

One of the key recommendations from the JISC UK LOCKSS Pilot Programme evaluation report was that steps should be taken to minimise the risks associated with UK based support. Support proved to be central to the overall success of the pilot. In considering the transition from pilot to service HATII at the University of Glasgow considered with the Digital Curation Centre its own mission and how this related to the rollout of a LOCKSS service. HATII is a research-led institute and where it runs services these have tended to be as part of research into technical, organisation, and structural aspects of such endeavours. This was the case in the UK LOCKSS Pilot Programme, we were interested to determine whether it was possible to implement an effective technical support service for the thirty-two participating institutions of the UK LOCKSS Pilot Alliance, whether we could construct a substantial collection of e-journals to which the participating institutions would have archival rights, whether we could raise the levels of community engagement with the LOCKSS initiative, and whether we could create the foundation for a self-sustaining UK alliance that will enable institutions to commit to the use of LOCKSS as an e-journal archiving solution following the end of the Pilot Programme. We succeeded in achieving each of the first three goals and believe that we have also achieved the fourth, but will only know for sure about this if the UK LOCKSS Alliance takes off. Despite the praise which the JISCs independent reviewers, Dalton and Conyers (2008) had for HATII's role in this initiative we took the decision that the rollout of a national service did not correspond to HATII's core mission.

Following discussions with EDINA at the University of Edinburgh and a collaborator in the Digital Curation Centre we took the decision to recommend to the JISC that for the development of the UK LOCKSS Alliance the LOCKSS Technical Support Service move to EDINA. The JISC accepted our recommendation and agreed that the EDINA mission to *“enhance the productivity of research, learning and teaching across all universities, research institutes and colleges in the UK by delivering first-rate online services and by working with support staff in university and colleges and with other partners in the academic community, and beyond, and by carrying out successful R&D projects”*¹³ was closely aligned with the objectives of the UK LOCKSS Alliance. LOCKSS will complement EDINA's growing set of electronic journal archiving related projects. For example, EDINA has been participating in CLOCKSS for over two years. Bringing LOCKSS and CLOCKSS together will ensure the two initiatives can work together to address the needs of the UK community and the full spectrum of relevant UK electronic journal content. EDINA has every likelihood of emerging as the national centre in the UK with expertise in journal archiving.

¹³ EDINA Website, <http://edina.ac.uk/about/>

Throughout the UK LOCKSS Pilot Programme, we explored the proposed methods by which libraries can access the content stored within a LOCKSS box. By design LOCKSS, itself effectively a transparent HTTP proxy server, was designed to integrate with an institutional proxy server. This would mean that when a client or web browser requested content available at some URL, the institutional proxy server would forward the request to LOCKSS box, which would in turn forward the request to the original publisher and only serve locally preserved content if the requested content was no longer available from the original publisher. However, participating librarians were not keen on this approach. Some institutions did not have an institutional proxy; others were hesitant to integrate LOCKSS into their overall institutional network environment during the pilot. Some questioned who would then be responsible for LOCKSS in the event of system failure; the network team, librarians, or LOCKSS support. There was an overwhelming preference for LOCKSS to serve content corresponding to OpenURLs, links specific to the LOCKSS system that could then be integrated into existing library-based link resolver systems. As a direct output of this discussion, the US-led development team has undertaken development work and a first implementation of the alternative mechanism was released at the end of July 2008. Moving forwards, the processes by which archived journals will be served to users need to be explored in greater detail and continue to be refined in response to experience. For a variety of reasons (e.g. in consideration of access problems by remote readers not able to access content in LOCKSS only available locally), readers should be made aware they are accessing archived material rather than that from the original publisher's website.

The Dalton, P. and Conyers, A. (2008) evaluation report indicated that publisher workflows needed to be improved. Publisher's needed more support on manifest page development¹⁴, perhaps a greater overview of the technology itself, and the situations in which the archived content would be accessed. Complexities have arisen because each publisher, and publisher platform, works in a slightly different way. Developing a generic walk-through that is useful for all, and yet does not confuse readers, has been challenging. We would be open to suggestions for mechanisms that in retrospect might have simplified the process. As we emphasise participation of small, medium and open access publishers the diversity of publisher platforms encountered is likely to increase. Currently, the process for releasing content in the LOCKSS system is complex and requires an involved quality assurance process to be followed. To increase the quantity of content that can be processed and released, we expect to explore the contributions libraries could make to this activity, fostering further knowledge and development effort in the UK community. This would, in addition, reduce

¹⁴ Manifest pages are the online pages hosted on a publisher's website that authorise an institution to collect and archive a journal volume through LOCKSS. They are only available to those institutions that have archival rights to the content.

dependencies on individual staff members that inevitably produce bottlenecks.

Conclusions

In this paper, we reflected on the process of running the UK LOCKSS Pilot Programme alongside the conclusions reached by the JISC commissioned evaluation reports, and looked at how we can move forward in the UK with LOCKSS-based archiving services. We assert that digital journal archiving can be considered a risk management activity and the UK community must collectively act to distribute and manage the risks associated with long-term access to electronic journals. As we have employed a risk identification approach and here we aim to highlight several pressing issues facing the UK HE/FE library community in long-term electronic journal archiving.

- While it is evident that libraries must actively take measures to prevent loss of access to digital content, it is not evident that one journal archiving approach is technically, culturally, economically or organizationally the best. Currently journal archiving benefits from the use of a variety of approaches.
- Librarians (and indeed publishers) will need a greater awareness of the risks and benefits associated with the different approaches to journal archiving and to factor this knowledge into their decision making processes.
- Different libraries may have different requirements for the delivery of content to users and these individualized needs must be taken into account in the development of archiving services.
- Licensing arrangements and agreements remain a problematic area for long-term preservation initiatives. Librarians, publishers and agents will need to work harder to ensure that the agreements for preservation are negotiated to the mutual advantage of all parties.
- There is a general lack of clarity regarding roles and responsibilities at the institutional, national body, and journal preservation service levels and this is hindering progress towards delivering archiving journal archiving options and solutions.
- Costs remain a sticking point for the development of long-term preservation services. In particular it is hard to justify the costs associated with long-term preservation and to do so within the context of the actual range of services currently being offered. Journal archiving service providers must demonstrate sound financial sustainability and provide a transparent and positive cost benefit ratio to their participating libraries.

From the outset of the Pilot in March 2006 it was evident that the challenges to journal preservation were not merely technical in nature but required that organisational, cultural, and structural challenges be reviewed and addressed. Participatory, collaborative, and distributed initiatives for preservation show real promise, and combining the technical strength of LOCKSS with the ability of the LOCKSS Alliance to

promote relationships with publishers and community driven action by libraries is very promising.

So building on the lessons learned from the UK LOCKSS Pilot Programme we are focused on establishing a stable and sustainable UK LOCKSS Alliance. Central UK coordination has proved valuable by ensuring UK specific issues are effectively identified and resolved consistently and at national level. Indeed having the JISC strongly backing the UK LOCKSS initiatives has been a very positive factor in ensuring their success. We feel that by bringing together institutions to share experiences we are facilitating the development within the information management and library communities of the concepts and issues surrounding journal archiving. As the programme enters the second phase UK LOCKSS will explore new ways in which libraries may contribute to developing journal archiving strategies and mitigate the inherent risks. In response to the concerns of librarians, publishers are increasingly participating in efforts to develop effective journal archiving strategies. By leveraging the skills of the community and integrating the library as an essential component of journal archiving, the UK LOCKSS Programme ensures that the key stakeholders affected by the challenges of the current environment are given appropriate opportunities to participate in the solution.

Acknowledgements

The UK LOCKSS Pilot Programme, which ran from 1 March 2006 through 31 July 2008, was jointly funded by JISC and CURL as part of the Digital Preservation and Records Management Strand (2006). The authors wish to thank colleagues at HATII at the University of Glasgow, the Digital Curation Centre (DCC), Paul Harwood, Albert Prior and Carolyn Alderson at Content Complete Ltd (CCL), Tony Kidd, William Nixon and Laura Roy at the University of Glasgow Library who ran the OpenLOCKSS Project Team, the JISC and JISC Collections teams (and in particular Neil Grindley and Lorraine Estelle), and the team at the US LOCKSS Alliance based at Stanford University. We extend special thanks to Vicky Reich and David Rosenthal who worked very closely with the UK LOCKSS Technical Support Officer. We also wish to thank the librarians and their system support staff at the thirty-two institutions which participated in the pilot programme: University of St Andrews, University of Birmingham, University of Bristol, Cambridge University Library, Cardiff University, De Montfort University, University of Durham, University of East London, University of Edinburgh, University of Exeter, University of Glasgow, University of Hertfordshire, University of Hull, Kings' College London, University of Leicester, University of Liverpool, Loughborough University, University of Manchester, Middlesex University, University of Newcastle Upon Tyne, Oxford University, University of Sheffield, University of Surrey, University of Sussex, London School of Economics and Political Science, University of Huddersfield, UCL Library Services,

University of Warwick, University of Wolverhampton, and University of York.

As we completed the final version of this paper (August 2008) Adam Rusbridge moved to EDINA at the University of Edinburgh where he will continue his links with the UK LOCKSS Alliance.

References

Hockx-Yu, H. 2006. Establishing a UK LOCKSS Pilot Programme. *Serials, Volume 19, Number 1*. <http://uksg.metapress.com/openurl.asp?genre=article&issn=0953-0460&volume=19&issue=1&spage=47>

Rusbridge, A., and Ross, S., 2007. The UK LOCKSS Pilot Programme: A Perspective from the LOCKSS Technical Support Service. *International Journal of Digital Curation, Volume 2, Number 2*. <http://www.ijdc.net/ijdc/article/view/49>

Morrow, T., Beagrie, N., Jones, M., and Chruszcz, J., 2008. A Comparative Study of e-Journal Archiving Solutions. http://www.jisc-collections.ac.uk/media/documents/jisc_collections/reports/e_journals_archiving_%20solutions_report_final_080518.pdf

Dalton, P., and Conyers, A., 2008. Evaluation of the JISC UK LOCKSS Pilot. <http://www.jisc.ac.uk/media/documents/programmes/preservation/uklockssevaluation.pdf>

Rosenthal, D., Robertson, T., Lipkis, T., Reich, V., Morabito, S., 2005. Requirements for Digital Preservation Systems: A Bottom Up Approach. *D-Lib Magazine, November 2005*. <http://www.dlib.org/dlib/november05/rosenthal/11rosenthal.html>

The KB e-Depot in development

Integrating research results in the library organisation

Hilde van Wijngaarden, Frank Houtman, Marcel Ras

hilde.vanwijngaarden@kb.nl, frank.houtman@kb.nl, marcel.ras@kb.nl

National Library of the Netherlands

Prins Willem-Alexanderhof 5

2509 LK The Hague

The Netherlands

Abstract

The mission of the KB e-Depot is to ensure permanent access to large quantities of digital resources in a national and international context. Operating an international e-journal archive at a relatively small organization such as the KB asks for a firm foundation of its policy. With support from the Dutch government, the KB has succeeded in setting up two expert teams: an operational team responsible for daily operations of the e-Depot (based in the Acquisitions & Processing Division) and an active research team that secures continuing research and development to secure long-term preservation and perpetual access to electronic information (based in the Research & Development Division). The experience gained in operating an archive is directly used in system- and process improvements. As are the results of the projects in which the R&D team is involved. The organisation around the e-Depot is based on this pragmatic approach.

National Library of The Netherlands

The Koninklijke Bibliotheek (KB, National Library of the Netherlands) is a scientific library and a deposit library. Its mission is to collect published information, preserve it and provide permanent access to the information for use in research, education or for any other purpose in society. In most countries, publications have to be deposited by law. The Netherlands does not have an act or provisions in law concerning depositing. KB works with a voluntary deposit system based upon agreements with the publishers. This has resulted in nearly complete coverage of the print publications produced by commercial publishers in The Netherlands.

Digital archiving system

In the early nineties of the last century, KB began discussing archiving of digital publications. In 1996 an agreement was signed with Elsevier and the first experiments with digital archiving started. The Dutch Publishers Association agreed on a new arrangement in

1999, which covered also online digital publications with Dutch imprint. The traditional model, based on national deposits and geographical boundaries, is no longer valid for guaranteeing the long-term preservation of the international digital academic output. Academic literature is produced by multinational publishers, and has often no longer a country of origin that can easily be identified. In line with the international nature of information provision, the KB decided to open up its e-Depot to international publishers in 2002.

In that year, a landmark archiving agreement with Elsevier was signed, including all Elsevier e-journals instead of the e-journals with Dutch imprint. This arrangement turned the National Library into the first digital archive in the world for e-journals published by international scientific publishers. Other publishers followed and currently, KB has agreements with 14 of the most important international publishers. In 2008, 11 million digital objects are stored in the e-Depot, mainly e-journal articles in pdf formats.

The core of the e-Depot is the Digital Information Archiving System (DIAS), developed during a two year project between 2000 and 2002. DIAS is a combination of standard IBM components, with extra functionality that allows the system to interact with the library infrastructure. [1]

Based on the agreements with the publishers, e-Journals are delivered to the KB and ingested into the system automatically. The error recovery procedure is the only manual effort involved. Metadata are delivered by the publisher and converted to the KB format and added to the KB catalogue. Access policies depend on agreements with the publishers. Commercial content can be accessed on site only and trigger events that will allow the KB to open up online are being discussed.[2]

In addition to the e-journal articles, more collections and different types of material will be stored in the e-Depot in the very near future. These new types of material will be more complex, but also be more voluminous.

Apart from the core DIAS, the e-Depot consists of different modules that allow pre processing and access. Especially these components are subject to improvements in the case of adding new types of material.

The national e-Depot

As mentioned above, procedures and workflow were initially designed for the national e-journal archive. But because of the nature of e-journal publications these were quickly dedicated towards an international e-journal context.

A selection of Dutch e-journals, deposited in the context of the KB's depository task, is ingested using the same procedures. That means e-journals of larger publishers, delivered in large quantities.

The national e-Depot is the digital version of the deposit of Dutch printed publications. Because of the broad variety of digital objects, acquisition and processing of these materials will be extended gradually. The first step was to set up digital archiving workflows for national deposit of singular e-journals and monographs. Therefore, web-interfaces were set up to allow any depositor to submit publications, monographs or periodicals. Also procedures were designed to process these singular publications. To allow the system and the organisation to process a growing number of very diverse materials, a number of changes and additions have to be implemented in the e-Depot infrastructure. The next step will be to set up workflows for complex objects like websites and other multimedia objects.

The archiving of digital deposit materials will bring changes in the organisation, especially concerning acquisition and processing of these publications. But also on the technical level extensions to the e-Depot system are needed to handle these new processes.

The Dutch web archive

Of a more complex nature are web sites. In 2005, KB started a project with the goal of harvesting and preserving a selection of Dutch web sites. As for the harvesting of web sites, but also the access to the archived web sites, it was decided to make use of standard tools used by other web archiving projects. By using the toolset developed under the colours of the International Internet Preservation Consortium (IIPC), the KB was able to focus mainly on the preservation issues concerning web sites. A selective approach was chosen because of the large volume of the Dutch web and the fact that it was decided to collect web site in full depth as part of the national deposit. A so called domain approach does not make sense in this case.

The project is concentrated on preservation issues, quality assurance and on setting up a procedure for processing web sites based on the current possibilities within the e-Depot.

To put this new procedure in production, a number of changes to the e-Depot infrastructures, as well as additions to the metadata model, are necessary to ensure the durability of the content. [3]

Digitised materials

Another new type of materials to be processed are the masters resulting from national digitisation projects carried out by the KB, e.g. Dutch newspapers from 1618 and Parliamentary Papers. These materials are well structured, but in other file types as we are used to process, Tiff and Jpeg200.

KB e-Depot phase 2

Operating a digital archive entails continuous efforts in improving workflow and infrastructure, maintenance. But also investing in research activities to develop tools that enable us to really do what the e-Depot was meant for: retrieving archived documents for eternity.

During the first six years the e-Depot is operating, the focus was on processing e-journal articles. Generally these are objects of a same type and form. Over 11 million e-journal articles have passed the processes since 2003. The initial workflow did work for the past six years. However, this same set-up will not be sufficient for the many different types of content and content-suppliers that are on the e-Depot itinerary for the next few years. Currently we work with a pre process which has only basic functionalities and only basic quality control on the materials to be ingested. That will have to be improved and could be improved on the basis of our own research and developments, but also based on the results of the PLANETS projects as described above.

Apart from the steady growth of the e-journal archive, the Dutch digital deposit, the web archive and the KB's active digitisation policy and the goal to preserve digital images are the main drivers for extensions and improvements of the current e-Depot. Therefore processing and storage capacity needs to be scaled-up enormously to be able ingest of new collections which tend to very voluminous.

During the last six years, KB also invested heavily in research. Preservation research delivered the initial design of preservation functionalities that are now ready to be implemented into the e-Depot infrastructure: characterisation tools, improvements to the pre ingest process, a normalisation module, a migration module and new requirements for the metadata model. Tools that are strongly needed to enable the improvement of the quality of processes and the procedures for quality control of the objects to be ingested.

To coordinate implementation projects of new functionality running at the same time as the enlargement

of the system and to control dependencies between these projects, KB has set up a programme. Projects run simultaneously while using each others' input. The programme is now in full speed and will deliver the e-Depot infrastructure 'phase 2' in the middle of 2009.

Operating a digital archive in a library

During the development of the e-Depot system in 2002, a working group was started to develop organizational embedding of the system. It was decided to make the Acquisitions & Processing Division responsible for day-to-day operations and to set up a digital preservation research team within the Research & Development Division. The IT department took on the daily maintenance of the system, with IBM staying closely involved. In January 2003, five people started their new responsibilities in the three departments. Today, six years later, 23 fulltime equivalents are dedicated to the e-Depot.

Still, the e-Depot department is responsible for daily operations. Seven collection managers and one functional manager perform tasks which are focused on processing of objects. These are subdivided into different specialist tasks based on the workflow. The Front Desk is responsible for technical contacts with publishers, analysis of content and metadata, guidelines and the set up of processes. The Pre Ingest group is responsible for the technical set up of the process, which means conversions of metadata, writing scripts and style sheets and quality assurance. The Ingest group is in charge of the actual ingest of materials into the DIAS system and error control.

The Digital Preservation department is responsible for preservation research and development. Their daily activities are directly related to the operational e-Depot while they are involved in different European R&D projects like Driver, KEEP, Parse.Insight and PLANETS.

The IT department is entrusted with the technical maintenance of the system and with the coordination of technical improvements on the system. Most of the work on these improvements is outsourced.

The group of KB-staff that has something to do with the e-Depot is even much bigger than that. Access is the responsibility of the User Services Division, management of the relations with publishers is done by the Acquisitions department and cataloguing by the Cataloguing department. All these departments are closely involved in the e-Depot as well.

We are now in a position to evaluate the consequences of running an operational digital archive for the library as a whole and move on to the next phase of improving workflow, enhancing the system and the quality assurance and take a major step in scaling up our storage- and processing capacity, as mentioned above. This could only

be possible because of the firm embedment of functions within the different places of the organization. Commitment of the library as a whole is vital for this.

The e-Depot is a driving force for renewal and change within the organization. The influence is noticeable in three areas. First, there is the content of the library's collections. Taking up the responsibility for long-term digital preservation has made the KB's collection more international, scientific and more diverse in appearance, now also containing multi-media applications, websites, e-books etc.

Secondly, substantial changes had to be implemented in the technological infrastructure that is now also beneficial to the development of other library services. This also includes changes in metadata modelling and handling.

Thirdly, there is the impact of running the e-Depot on people and the organisation. To organise digital preservation activities across several departments is not an obvious choice. And it has not always been the easiest choice either. It requires special attention to coordinate between different departments and to set-up good knowledge management and quality assurance.

However, after six years, we can say it has been worth it. The digital preservation research team could focus on research issues and set up an active role in international projects, but with a firm practical basis and focus on implementable solutions.

The e-Depot department was a separate team in the Acquisitions and Processing division at the start, but is now growing and becoming more and more interlinked with the rest of the division. While all processes are becoming digital (eg. automatic metadata ingest and processing for printed publications), differences between the digital depot and the traditional depot are becoming smaller. The best example of this integration of separate processes is the automatic handling of publisher-submitted metadata. The e-Depot has been working with submitted metadata since the start, while metadata for the print collections is generated manually at the library (in case the metadata is not yet provided by others in the shared national catalogue). For some collections, KB is now developing import of metadata records, setting up a similar workflow for processing of both print- as digital collections.

The e-Depot department with its staff working with the newest digital procedures in an international environment now co-operates closely together with library staff with 'traditional' library skills in the area of acquisition and cataloguing. The e-Depot meant doing the same kind of work in a completely different way and brought people to the library with a new set of skills. And through the

interchange within and across divisions, people can learn from each other and get familiar with new digital processes.

But running the e-Depot at the KB also brought liveliness, an international atmosphere and a broader outlook on the information landscape, thus making the library a really attractive place to work.

Preservation research

The success of the KB e-Depot is built on two pillars: the operational digital archiving environment and a substantial investment in research. As mentioned before, the first practical results of the digital preservation R&D are now being implemented into the e-Depot infrastructure. This improves the quality of the system and the content to be ingested. Improvements are focused on the different phases of the process.

- delivery of objects
- workflow and management of workflow
- characterization of the objects to be stored
- collection management
- preservation management
- IT infrastructure

Implementation is organized in the program described above, that combines these new additions with the upscale of the loading and storage capacity of the system.

Newly developed preservation functionality is partly the result of extensive international collaboration. KB is an active participant in international projects to develop new tools and services. KB takes part in projects for two main reasons. First, KB is able to bring in Library specific knowledge and practical experience in digital archiving. Second, insight in new technologies is necessary to maintain the e-Depot infrastructure. The results of projects like Planets could be of direct use for the operational processes and infrastructure of the e-Depot.

Within Planets, the KB is responsible for leading the subproject Preservation Action.[4] Furthermore, the KB is participating in several other subprojects. As the Planets project is halfway now, it seems to be the right moment to evaluate the Planets output (in this case and more specifically the Preservation Action output) against the interest of an operational system like the e-Depot. Planets will deliver a sustainable framework to enable long-term preservation of digital content. Either the framework or the individual modules delivered by the project will be of direct use for the e-Depot.

PLANETS

Much has been written and said about the objectives of the Preservation and Long term Access through Networked

Services, or Planets project. In short, the main goal of Planets is to increase Europe's ability to ensure long term access to its cultural and scientific heritage. Planets delivers preservation planning functionality enabling organizations to plan their preservation actions in a structured and controlled manner. To characterize digital objects, Planets develops methodologies, tools and services, while preservation action tools will be in place to migrate or emulate digital objects. A testbed is created for the objective evaluation of different protocols, tools, services and complete preservation plans. The Interoperability Framework will integrate these tools and services in a network.

For the KB, Planets means performing the R&D we had planned, but in the setting of a closely collaborating international team. Requirements for Planets tools and services are based on KB's practical experiences and future plans, but also aimed at developing a more general framework. Planets products should not be specific for one organization, but should offer a set of services for a large variety of institutions. In practice, for the KB as participant in Planets, this can cause some tension because resources go into developments that might not be directly implementable in the KB. At the same time these activities are necessary to create the overall framework.

The project Planets being two years on the way, we want to take a look at some of the project results and will evaluate what these products could mean for the further development of our e-Depot environment. Within the scope of this paper we restrict ourselves to products that are/will be delivered by the subproject Preservation Action only. This subproject is concerned with the creation of solutions to perform preservation actions. In other words, this subproject is responsible for making the tools available that are needed for rendering digital objects, either in a different format (migration tools), or in a different technical environment (emulation tools). Next to migration and emulation tools the subproject also includes the development of a Tool Registry and a variety of reports of a more strategic purpose. In the following we will discuss the products delivered by the subproject and the possible value for the e-Depot environment. Subsequently we will discuss the Preservation Action Blueprint, the GAP Analysis, the Tool Registry, the Preservation Action Tools on Emulation and Migration.

The Preservation Action Blueprint

One of the products of the PA sub-project is a Blueprint that can be used by any developer or supplier when developing new preservation action tools, - both migration and emulation. It provides a list of functional requirements that these types of tools should offer. It also presents the workflow that should be followed when incorporating newly developed tools into the Planets framework. This list of functional requirements for newly build, improved or

adapted PA tools will ensure not only a consistent behaviour but a consistent level of quality as well.

Since the Blueprint is very much aimed at guiding and stimulating future development, for now it is just of indirect value to the e-depot workflow. In future, it will guarantee a certain degree of quality of new PA tools. It can (and should...) be used by developers when building new preservation action tools.

GAP Analysis

Within the Planets project, the PA subproject is responsible for providing tools required to perform preservation actions. In order to do so, existing tools can be wrapped and made available within the Planets framework. If no tool for a certain action exists, new tools have to be provided for. To offer a choice of tools to be used for preservation actions, we first need to know which file formats are in use for long term archiving. This is what has been done in the project: we created a list of file formats based on information provided by 76 institutes from different countries. At this moment (August 2008) the list contains 121 used file formats but is still being expanded. By analyzing this inventory we will have a clear understanding of which preservation action exist and/or what tools are needed.

What we have found for the Blueprint is also true for the Gap analysis. There is a certain value for the e-depot, but again it is of indirect value for the e-Depot although it could constitute an important instrument in the future when combined with the tool registry.

Registry

The Planets Preservation Action Registry stores descriptive information about preservation action tools (and services, which are wrapped tools) and how and for what kind of actions to use them. In Planets PA registry, a preservation action tool is a software program that performs a specific action on a digital object to ensure the continued accessibility of this digital object. This action could result in a transformation of the object or a (re)creation of the technical environment required for rendering the object, or result in a combination of these two.

How tools and services could be used is described in a 'pathway'. A pathway is a predefined set of one or more preservation actions operating on a specific input file format and version and possibly (in the case of an 'actions on objects' tool) resulting in a specified output format. A pathway can include at least one or more preservation actions (and thus require at least one or more tools).

Of course, a registry which includes indications of both functionality and quality of preservation action tools contains a very usable overview for the e-Depot. The

registry will be of direct use to deploy preservation action tools before or after ingest.

Migration tools

As more and more heterogeneous content will be presented to the e-Depot (think for example about the content that is generated by web archiving), the need for preservation actions becomes more important. One of the main digital preservation strategies is migration. Migration modifies a digital object in order to keep it accessible. There are three types of migration to distinguish. First, there is a type of migration that will take place in the ingest phase. This we call normalization. At the moment, a module for the e-Depot is in development that will convert text based publications that are not delivered in the PDF format, to PDF/A. There is a second type of migration that will be periodically used to execute batch migrations. This kind of migration will be used to prevent already stored digital objects to become obsolete. The third type is called migration on demand. A digital object will be temporarily migrated at the request of a user.

Migration tools, or tools for objects, are essential in the e-Depot environment. They play an important role in the pre-ingest phase, during the storage phase and eventually in the access phase. The results of the Planets project will offer a broader range of migration tools that will allow a digital archive like the e-Depot to perform migrations of a higher quality, including quality control.

Emulation tools

Another strategy to ensure the accessibility of digital documents is formed by emulation. Emulation tools, or tools for environments, change the technical environment in such a way that the original objects can be accessed. In Planets, a modular emulator is developed, based on earlier research and development of the KB. This emulator, named Dioscuri, is especially designed for digital preservation by being more durable and flexible than other emulators.[5] In the design, each hardware component is represented as a module in the emulator. A full emulator is created by combining all the modules.

Emulation tools could play a significant role in the accessibility of digital objects. For KB, emulation is as important as migration, because a growing group of digital collections (interactive, complex objects) cannot be migrated if the original does not work anymore due to their complexity. The development of the emulation tool within Planets is again an investment in the future.

Employability

Several Preservation Action products can be, directly or indirectly, employed within the e-Depot workflow. Obviously, within the Planets project many more significant tools and products with a potential value for the workflow in the e-Depot are developed. For example, in

the Preservation Planning part of Planets the planning tool PLATO is developed, while the characterization module PRONOM will be further developed in the Characterization subproject. However, within the scope of this paper it is impossible to describe all the (potential) valuable tools and products at some length.

Conclusions: Workflow improvement

Six year of operating a digital archive brought a lot of practical experience and knowledge. At the same time years of research started to pay off. The KB R&D department delivered a clear list of functionalities that now have to be implemented to ensure ongoing durability of ingested material. Because of this, the KB has a firm fundament to raise the level of the e-Depot environment and to extend the usability.

People from different departments (e-Depot, Digital Preservation and IT) are now in a program together to implement changes. And at the same time, research is moving forward, working on rendering tools like the emulator Dioscuri, which will be tested on the results of the KB's web archiving project.

The results of the Planets project as described above generate new tools which can be directly implemented in the e-Depot environment. Besides, these research activities create new views on preservation issues and the workflow of KB's e-Depot environment. Practical experience and knowledge from research projects give us a clear focus for further research. Current developments with the program setting up a new ingest process make us feel confident about the phase after that: when current research delivers results and will become implementable solutions for permanent accessibility.

References

- [1] Erik Oltmans and Hilde van Wijngaarden, Digital Preservation in Practice: The e-Depot at the Koninklijke Bibliotheek', in: *VINE - The Journal of Information and Knowledge Management Systems*, Vol. 34 (1), pp. 21-26.
- [2] Erik Oltmans and Adriaan Lemmen, The e-Depot at the National Library of the Netherlands. In: *Serials*, Vol. 19 (1), 2006, p. 63-67 and Els van Eijck van Heslinga., 'SHAPING COURSE. The Development of the Strategy for the e-Depot of the Koninklijke Bibliotheek, National Library of the Netherlands, in a National and International Context' In: *New Technology of Library and Information Service (iPres2007)*, 2007.
- [3] Information on the KB webarchive is to be found at: http://www.kb.nl/hrd/dd/dd_projecten/webarchivering/index-en.html
- [4] The Planets website can be accessed at: <http://www.planets-project.eu/>. Information KB in Planets:

http://www.kb.nl/hrd/dd/dd_projecten/projecten_planets-en.html

- [5] Jeffrey van der Hoeven, Bram Lohman en Remco Verdegem, Emulation for Digital Preservation in Practice: The Results, *International Journal on Digital Curation (IJDC)*, Vol. 2 (2), 2007.

Building a digital repository: a practical implementation

Filip Boudrez

Consultant digital archives

City Archives of Antwerp

Oudeleeuwenrui 29

2000 Antwerp

Belgium

filip.boudrez@stad.antwerpen.be

Abstract

Implementation of a digital repository involves much more than just the installation of the required hardware and software. It is clear from the implementation path undertaken by the Antwerp City Archives over the past year and a half that it is also an organisational issue and requires fundamental consideration of core records management and recordkeeping issues.

Introduction

Antwerp City Archives began the implementation of their digital archive repository in Autumn of 2007. Such a project however, involves much more than simply installing the required hardware and software. It is not only a matter of many organisational issues; the construction of a digital repository also raises fundamental questions about records management and recordkeeping.

1. The digital repository

Digital archiving has, for a long time, been a policy priority for the Antwerp City Archives. A proactive policy has already been in place for many years, focussing on city agencies and services that create records. During this time, the transfer of digital records to the archival service began and also many archival documents were digitised internally. Thus the volume of born-digital and digitised archives managed by the City Archives has gradually been on the increase.

The Antwerp City Archives anticipates that the digital repository will address procedures and infrastructure for the ingest, the management, and the dissemination of the digital archives and collections with medium to long term retention periods. The digital resources archived in the digital repository must be authentic and durable, regardless of whether they are born digital or digitised.

Construction of the Antwerp City Archives digital repository has, to a large extent, been based on the

research and recommendations from eDAVID.¹ Development and programming has, for the majority, been carried out in-house. Only the dissemination has been contracted out to an external developer, because this process is carried out separately through the website of the Antwerp City Archives.

2. An integrated recordkeeping system

An important and central aspect of the implementation is the positioning of the digital repository within the city archives. It is widely accepted that there are two main options for implementing this. One option is for the digital repository to be constructed alongside the recordkeeping system for analogue records and archives. This option results in only minimal integration between paper and digital records. The only integration envisaged in such instances is the use of metadata, to make the collections easier to search by researchers and civil servants working in the city administration. An alternative option is complete integration of paper and digital recordkeeping. In this case, the same procedures are used for treating and managing both paper and digital archives.

The City Archives has opted to implement option two for their digital repository and is therefore seeking maximum integration between paper and digital archives. There are many reasons for this. The first is so that the digital repository of the Antwerp City Archives may not be an isolated system. The second is that the City Archives wishes to implement a single system that will cater for all archives and collections, regardless of the medium or form of the records. The advantage of this is that the archival processes for intake, management and giving access are consistent for all types of records and only need to be automated once. The same software system can be used for managing all archives so that, in as far as is possible, information is shared and only needs

¹ See <http://www.edavid.be/eng/index.php>.

to be registered once. The principle of authority records is thus applied in a de facto manner and the need for developing similar functionality (for example, the ability to construct detailed retrieval aids) across different systems is avoided. Thirdly, integrated management of the archives is completely in line with the records creation practices that exist in a hybrid records creating environment. Finally, integration also contributes to embedding the digital repository within the institution.

The Antwerp City Archives is therefore constructing a single integrated recordkeeping system in which the management of paper and electronic records is fully part of. The integration takes place across numerous levels, including the software for processing and managing the records, the information architecture and the metadata model, and finally also the procedures.

3. Recordkeeping software: MACZ

Significant opportunities existed at the start of the implementation project that impacted on the choice of software for archives and records management. With the move to the Sint Felixpakhuis and the denBell project it was recognised that records management within the Antwerp City Archives was in need of revision. The new location and the application of box placement according to their size meant that a thorough and automated repository system was needed. Prior to the move, work also commenced on structuring and automating physical management practices in the archives.²

The denBell project started very soon after the move. This will soon result in administrative departments transferring large volumes of records to the archival service at pace. This must be done in an efficient way so they are quickly processed and accessible after their transfer. To facilitate this and ensure it is most effectively achieved, the way in which records are inventoried has been completely revised and the three international standards for archival description, ISAD(G), ISAAR(cpf) and ISDF, have been implemented. The way in which transfer lists were composed and processed has also been assessed and entirely adapted.

The new recordkeeping system has been named MACZ. MACZ will be completely managed and developed internally. As a result of opting for an integrated system, it was clear from the outset that the digital repository would be involved in the development of MACZ. The

archives will not therefore be managed by multiple and different information systems.

4. Information architecture and the metadata model

In revising the information architecture and data model for recordkeeping, paper-based records management was not necessarily assumed to deliver the leading principles. This wouldn't be strange as paper records management was a very familiar process. Instead, everything was thoroughly evaluated and optimum functional records management approaches and solutions were sought. Eventually it became apparent that the basic principles of digital archiving would provide the starting point for the new integrated information architecture and the new metadata model.

Two central principles for the new recordkeeping system in the Antwerp City Archives are, on the one hand, the separation of physical records management aspects from intellectual ones and, on the other hand, approaching archival records as abstract entities with one or more representations (paper, digital, microfilm/fiche etc). Separating physical and intellectual records management aspects has however a few consequences. The two aspects are managed and described separately but must remain clearly identifiable. Moreover, they must continue to be clearly linked to one another. In practice, this is achieved by linking the inventory numbers on the ISAD(G) records with descriptions of the records.

Another consequence was that ISAD(G) could not be implemented on its own and that the ISAD(G) implementation at the Antwerp City Archives deviated from the standard on numerous points. Diverse ISAD(G) description fields are related, for example, to representations of the archival documents (e.g. the element 'Extent and medium of the unit of description') and would be better described on the level of the inventory number. Such a ISAD(G) fields were not consequently implemented as proscribed by the standard but were given new functions.

This approach of identifying archival records using inventory numbers was already in place. This manner of working will be extended from paper record keeping to the digital archives. An inventory number is assigned to archival files and archival items as logical entities. For digital archiving a small amendment will be made: a single inventory number can refer to an entire series of digital archives to facilitate the complete retrieval and consultation of the full series. Request and retrieval of a digital series therefore has no logistical restriction in the way that paper archives do. In fact, this means that for the identification of a digital archive series an inventory number will be used in the meaning of ISAD(G)'s 'reference code' element

² See the annual report of the Antwerp City Archives (2006 – 07) for more information. Annual reports are available at www.felixarchief.be. More information about the denBell project is also available from this site, though in Dutch only.

Different representations of the same archival record are not allocated individual inventory numbers. On the level of the inventory number itself is registered which representations of the same documents is in the holding of the City Archives; for example, archival records in paper form have the inventory number linked to the physical storage box number (one or more); digital records will have a reference to their location within in the digital repository; and records on microfilm will be denoted by the microfilm number (one or more).

The further allocation of inventory numbers and their extension to digital records had, for the city archives, the significant advantage that the archives staff already was already familiar with this method and did not require any additional training. The manner in which digital records would be prepared for ingest into the repository was communicated over internal email. The new recordkeeping software supports the above in different ways. Furthermore, a built-in functionality enables members of staff to examine for themselves the suitability of the proposed digital inventory numbers.

5. New record keeping procedures and development

An important consequence of pursuing a completely integrated records and archives management system is that the procedures for paper and digital records management must be fully aligned. The practical actions for processing paper and digital records are naturally different, but will be embedded in the same basic procedures and follow a common workflow.

The original intention here was to take existing procedures for paper records keeping as the starting point. During this exercise it became apparent that many procedures for paper records keeping were not consolidated and that analogue archives were sometimes processed by archives staff in very different ways. The implementation of a structured and extensively automated recordkeeping system meant that it was consequently necessary to first ensure that consensus was reached between staff and that procedures were agreed across the board. This took a lot of meetings and time, and for this reason, the implementation of the digitale repository slowed down.

In parallel with planning the new recordkeeping and procedures, it was also necessary to programme and test the required software modules for the digital repository. The components of the different OAIIS-processes – namely ingest, management, and dissemination - have been clearly defined and implemented. The functional model defined in the OAIIS standard results in a basic workflow but this does not have to be strictly followed whilst developing and implementing the different

software modules. Instead, the essential components of each step of the process are developed simultaneously. For example, as soon as the first digital inventory numbers are ingested in the digital repository, the essential management quality controls must already be operational and the digital records must be retrievable and accessible. The ingest process does not therefore have to be fully implemented before development and implementation of the management modules is started.

This approach means that after the groundwork has been done on essential modules, work can begin on further refinement and optimisation. As a result, archives staff can already begin learning and gaining experience with the system. One advantage of this internal development is the ease and speed with which feedback can be given. The same approach will also be (more or less) followed to develop the dissemination functionalities retrieval and giving access. Until now, this is the only part of the digital repository that has not been developed by the City Archives itself but has been contracted out to an external partner as front-end access to digital or digitised archives is integrated into the website of the City Archives. Development of the front-end portal is happening in two phases. During the first phase, the basic modules and interfaces will be developed. These will only be available to archives staff, who can retrieve and consult the digital archives. During the second phase, the remaining functionality will be implemented (registration of consultations, security etc) and will be further optimised according to feedback. The request process will be refined according to feedback from users and will be made more intuitive for the end user.

Gaining experience is valuable for both archival staff and external developers. A digital repository differs in numerous respects from other types of computing systems. Electronic records can have very complex structures (for example, a website), or can be extremely bulky (for example a digitised register). Such factors must be taken into account when the technical issues relating to query and consultation functionality are being worked out.

Alongside development of the first modules, the storage infrastructure was installed and configured. Digital objects will be securely and safely stored in a SAN (Storage Area Network). The available storage capacity of the digital repository will be systematically expanded on a step-by step basis. This step-by-step approach will safely support tests of future development activities to increase storage capacity. Additional and particular points of interest included definition of the security protocols, organisation of the back-up process, and the integration of the storage infrastructure. The storage environment must be secure not only against regular users but also applications such as those used within the front-end access portal. The digital repository differs from regular digital information systems, and this even extends to the backup systems. Due to the high volume

of backup data, a standard backup regime is insufficient. The creation of a full backup is a time consuming activity and cannot be carried out weekly or even twice a month. Nevertheless, the creation of good backups remains an essential part of a disaster recovery strategy.

Concern over this last point and a supporting risk analysis resulted in significant attention being paid to the arrangement of the storage infrastructure, which was not selected for technical or efficiency reasons but for providing support to meet minimum archival requirements/logic. Should data be lost at the database level or subsequent references become incorrect, the essential linkage and management data can quickly be generated anew or corrected.

6. Transferring the records into the digital repository

As soon as the essential modules for ingest and management were operational in the storage infrastructure, transfer of items into the digital repository could commence. By 2007, the Antwerp City Archives already had collected 1TB of digital and digitised records. These digital records had until then mainly been kept on CD, with some exceptions (f.i. preserved websites) stored on file servers. Transferring these legacy files was thus one of the first tasks.

As a result of the sheer amount of work and the additional activities this task entailed, it was something of a sub-project in its own right. Just copying the contents of the CDs had already taken a long time. Moreover, the contents of the CDs could not simply be transferred direct into the digital repository. In most cases, one or more archival processes had to be undertaken upon the digital records: assigning inventory numbers, abolishing inventory numbers not accepted by the new system and therefore neither by the digital repository, adapting folder names, checking integrity, registering metadata that had been distributed up until then, and so forth. For the most part, these archival processes generally required additional and meaningful checks or analysis so dealing with them took a considerable number of months.

As soon as the first digital inventory numbers were ingested in the digital repository, essential management tasks had to be carried out. This assumed in the first instance that responsibility for management of the digital repository had already been clearly determined. It was expected that the digital repository manager would be thoroughly acquainted with the metadata and database models. This was already the case for the Antwerp City Archives, for not all of the required user interfaces and associated modules for managing the system were ready and a certain amount of work still took place behind the scenes.

A particular activity in this part of the implementation process involved the AVA image database. AVA contains descriptions of digital photographs and makes them available online. Master copies are now stored in the digital repository, whilst low resolution copies are uploaded to AVA for access. The digitised photos are therefore managed in both systems, which means AVA must be integrated with MACZ and the digital repository. However, as AVA was developed externally it has not been possible to integrate all of the functionality in with the recordkeeping system. Eventually it became clear that the most important integration work would be realised within the MACZ environment. The AVA database itself needed only one minimum adaptation. Once this adaptation was complete, work could commence to transfer the AVA master files into the digital repository.

Implementation of the digital repository has led to questions being asked about the future of the AVA database and its content. The digital repository enables direct access to digital photos via the City Archives' website without needing to access AVA itself. Numerous photos and pieces of iconographical work that had been previously stored in AVA are no longer stored there as these contents have been transferred to the digital repository. The AVA records will be deleted once the descriptions have also been extracted.

7. Conclusion

From the implementation path followed by the Antwerp City Archives, it is clear that the construction of a digital repository is initially more concerned with structured and efficient record keeping than with digital archiving in particular. Consequently, the implementation of a digital repository should ideally begin with a records management and recordkeeping 'vision' in which the best parts of paper and digital records management are brought together. Once the general procedures have been established for ingest, management, and making archives available, implementation of the digital repository should proceed relatively smoothly.

Implementation at the Antwerp City Archives has taken place in a step-by-step manner. Because of this it has been possible to learn about and systematically align records management practices, and thus also improve the digital repository. Implementation is therefore also an iterative process, in which consolidation and improvement continuously alternate with each other. The digital repository remains thus a work-in-progress.

Bit Preservation: A Solved Problem?

David S. H. Rosenthal
Stanford University Libraries, CA

Abstract

For years, discussions of digital preservation have routinely featured comments such as “bit preservation is a solved problem; the real issues are ...”. Indeed, current digital storage technologies are not just astoundingly cheap and capacious, they are astonishingly reliable. Unfortunately, these attributes drive a kind of “Parkinson’s Law” of storage, in which demands continually push beyond the capabilities of systems implementable at an affordable price.

This paper is in four parts:

- *Claims*, reviewing a typical claim of storage system reliability, showing that it provides no useful information for bit preservation purposes.
- *Theory*, proposing “bit half-life” as an initial measure of bit preservation performance, expressing bit preservation requirements in terms of it, and showing that the requirements being placed on bit preservation systems are so onerous that the experiments required to prove that a solution exists are not feasible.
- *Practice*, reviewing recent research into how well actual storage systems preserve bits, showing that they fail to meet the requirements by many orders of magnitude.
- *Policy*, suggesting ways of dealing with this unfortunate situation.

Introduction

For years, discussions of digital preservation have routinely featured comments such as “bit preservation is a solved problem; the real issues are ...”.¹ Indeed, current digital storage technologies are not merely astoundingly cheap and capacious, they are astonishingly reliable. Unfortunately, these attributes drive a kind of “Parkinson’s Law” (Parkinson 1957) of storage, in which demands continually push beyond the capabilities of systems implementable at an affordable price.

This paper is in four parts. The first part examines a typical claim made by a storage system vendor for the reliability of their product. It concludes that these numbers provide no useful information for bit preservation purposes.

Copyright ©2008 David S. H. Rosenthal

¹The prevalence of this meme is aptly illustrated by the letter from the programme committee accepting this paper. It cites the title as “Bit Preservation - A Problem Solved”.

The second, theoretical, part asks what characterizes a solution to the bit preservation problem adequate to the large numbers of bits to be stored and the long durations for which these bits are to be preserved. It proposes “bit half-life” as a metric for bit preservation, discusses the requirements being placed upon preservation systems in terms of this metric, and investigates the feasibility of benchmarking systems to see if they meet these requirements. It concludes that the requirements are so onerous that it is not feasible to measure whether systems meet them.

The third, practical, part reviews recent investigations into the performance of large-scale storage systems and their components. These studies uniformly report that storage reliability actually delivered to applications such as digital preservation systems is much less than that claimed by the manufacturers of systems and components. Tracking these failures to their root causes shows that every single hardware and software component contributes to some extent to the failures the systems experience. It concludes that current storage technologies fall well short of current requirements for bit preservation.

Given that the actual performance of storage systems is much worse than required, and that even if it improves we still won’t be sure that a system will meet its requirements, the fourth part asks what is to be done. As with paper, content in digital archives will inevitably suffer loss and damage. The question is how to invest the limited funds available for preservation to the best effect in terms of improved data survival. There are many ways in which spending more money can reduce (but never completely eliminate) the probability of loss and damage. What is needed to allow informed investment decisions? How can we encourage the development of cost-effective techniques for long-term bit preservation?

Clarification

It is incumbent on those attacking ideas such as the “solvedness” of bit preservation to focus on the strongest version of the idea². If proponents really believed that bit preservation was solved, they wouldn’t bother with backups. Of course,

²“we should always try to clarify and to strengthen our opponent’s position as much as possible before criticising him” (Popper 1959)

they do. What they really mean by bit preservation being solved is that the set of techniques in common use make it so unlikely that bits will be lost that there is no need for concern at the prospect.

The techniques in which they place such faith are backups and checksums. Their real belief is that if they make a few backup copies of their content, and include in them checksums which they occasionally verify, their content will be safe. The goal of this paper is to show that, while backups and checksums may be adequate for relatively short periods and small amounts of preserved data, the scale and duration of current preservation tasks render them inadequate.

The state of our knowledge about preserving bits can be summarized as:

- *The more copies the safer.* As the size of the data increases, the per-copy cost increases, reducing the number of backup copies that can be afforded.
- *The more independent the copies the safer.* As the size of the data increases, there are fewer storage options available. Thus the number of copies in the same storage technology increases, decreasing the average level of independence.
- *The more frequently the copies are audited the safer.* As the size of the data increases, the time and cost needed for each audit increases, reducing their frequency.

Thus techniques that might be adequate at a small scale will break down as the scale increases.

Claims

How would we know if bit preservation were a solved problem? I suggest that proponents of this claim must feel confident that they could at a minimum preserve a petabyte of data undamaged for a century. Petabyte-scale data collections with long-term value, such as the Sloan Digital Sky Survey (SDSS 2008) and the Protein Data Bank (WWPDB 2008) already exist, so this is asking them to surmount a rather low bar. How confident should proponents feel in their ability to keep a petabyte for a century? I suggest that they should have at least a 50% chance of success. Again, this is a rather low bar.

Proponents might bolster their case that these bars can easily be surmounted by pointing to claims such as: “ST5800 has a MTTDL (Mean Time To Data Loss) of 2.4×10^6 years.”³ (Sun Microsystems 2008), or: “a Pergamum system capable of storing 10^{16} bytes of user data [will have] an MTTDL of 1.25×10^7 hours, or about 1,400 years.” (Storer et al. 2008). These, and similar claims by other vendors, at first glance make it appear that bit preservation is indeed solved. Off-the-shelf solutions are ready to hand with performance so good that backups and checksums are quite superfluous. But do these claims stand up to examination?

Before using Sun’s claim for its ST5800 as an example, I should stipulate that the ST5800 is an excellent product.

³Numbers are expressed in powers-of-ten notation to help readers focus on the scale of the problems and the extraordinary level of reliability required.

It represents the state of the art in storage technology, and Sun’s marketing claims represent the state of the art in storage marketing. Nevertheless, Sun does not guarantee that data in the ST5800 will last 2.4×10^6 years. Sun’s terms and conditions explicitly disclaim any liability whatsoever for loss of, or damage to, the data the ST5800 stores (Sun Microsystems 2006) whenever it occurs.

All that the claim says is that if you watched a large number of ST5800 systems for a long time, recorded the time at which each of them first suffered a data loss, and then averaged these times, the result would be 2.4×10^6 years. Suppose Sun watched 10 ST5800s and noticed that three of them lost data during the first year, four of them lost data after 2.4×10^6 years, and the remaining three lost data after 4.8×10^6 years, they would be correct that the MTTDL was 2.4×10^6 years. But we would not consider that a system with a 30% chance of data loss in the first year had solved the bit preservation problem. A single MTTDL number isn’t a useful characterization of a solution.

Consider the slightly more scientific claim made at the recent launch of the SC5800 by the marketing department of Sirius Cybernetics⁴: “SC5800 has a MTTDL of $(2.4 \pm 0.4) \times 10^6$ years”. Sirius thus claims that about 2/3 of the failures occurred between 2.0×10^6 and 2.8×10^6 years after the start of the experiment. They didn’t start watching 10 SC5800s 2.8 million years ago. So how would they know?

Perhaps, instead of watching say 10 systems for 2.4×10^6 years they watched more systems for a shorter time. Sirius says they will sell 2×10^4 SC5800s per year at $\$5 \times 10^4$ each (a billion-a-year business), and they expect the product to be in the market for 10 years. The SC5800 has a service life of 10 years. So if Sirius watched their entire production of SC5800s ($\$10^{10}$ worth of storage systems) over their entire service life the experiment would end 20 years from now after accumulating about 2×10^6 system-years of data. If their claim is correct they would have about a 17% chance of seeing a single data loss event.

In other words, Sirius Cybernetics claims that the probability that *no SC5800 will ever lose any data* is over 80%. Or, since each SC5800 stores 5×10^{13} bytes, that there is an 80% probability that 10^{19} bytes of data will survive 10 years undamaged.

If one could believe the Sirius Cybernetics claim, the petabyte would look pretty safe for a century. But the claim clearly isn’t based on an experiment that won’t provide results until 2028 and even when it does will not validate the number in question. In fact, numbers like these are not the result of experiment at all. No feasible experiment could validate them. They are *projections*, based on models of how components of the system such as disks and software behave.

The state of the art in this kind of modeling is exemplified by the Pergamum project at UC Santa Cruz (Storer et al. 2008). Their model includes disk failures at rates derived from (Schroeder and Gibson 2007; Pinheiro, Weber, and Barroso 2007) and sector failures at rates derived from

⁴Purveyors of chatty doors, existential elevators and paranoid androids to the nobility and gentry of this galaxy (Adams 1978).

disk vendor specifications. Their system attempts to conserve power by spinning the disks down whenever possible; they make an allowance for the effect of doing so on disk lifetime but it isn't clear upon what they base this. They report that the simulations were difficult:

“This lack of data is due to the extremely high reliability of these configurations - the simulator modeled many failures, but so few caused data loss that the simulation ran very slowly. This behavior is precisely what we want from an archival storage system: it can gracefully handle many failure events without losing data. Even though we captured fewer data points for the triple inter-parity configuration, we believe the reported MTTDL is a reasonable approximation.”

Although the Pergamum team's effort to obtain “a reasonable approximation” to the MTTDL of their system is praiseworthy, there are a number of reasons to believe that it overestimates the reliability of the system in practice:

- The model draws its failures from exponential distributions. They thus assume that both disk and sector failures are uncorrelated, although all measurements of actual failures (Bairavasundaram et al. 2008; Talagala 1999) report significant correlations. Correlated failures greatly increase the probability of data loss (Baker et al. 2006; Elerath and Pecht 2007).
- Other than a small reduction in disk lifetime from each power-on event, they assume that failure rates observed in always-on disk usage translate to their mostly-off environment. A study (Williams et al. 2008) published after their paper reports a quantitative accelerated life test of data retention in almost-always-off disks. It shows that the 3.5" disks anticipated by the Pergamum team have data life dramatically worse in this usage mode than 2.5" disks using the same underlying technology.
- They assume that disk and sector failures are the only failures contributing to the system failures, although a study (Krioukov et al. 2008) shows that other hardware components contribute significantly.
- They assume that their software is bug-free, despite several studies of file and storage implementations (Jiang et al. 2008; Engler 2007; Prabhakaran et al. 2005) that uniformly report finding bugs capable of causing data loss in all systems studied.
- They also ignore all other threats to stored data (Rosenthal et al. 2005) as possible causes of data loss. Among these are operator error, insider abuse and external attack. Each of these has been the subject of anecdotal reports of actual loss of preserved data.

What can models like this tell us? Their results depend on both:

- the details of the simulation of the system being studied which, one hopes, accurately reflect its behavior, and
- the data used to drive the simulation which, one hopes, accurately reflect the behavior of the system's components.

Under certain conditions, it is reasonable to use these models to compare different storage system technologies. The most important condition is that the models of the two systems use the same data. A claim that modeling showed system *A* to be more reliable than system *B* when the data used to model system *A* had much lower failure rates for components such as disk drives would not be credible.

These models may well be the best tools available to evaluate different techniques for preventing data loss, but they aren't adequate to determine whether bit preservation is a solved problem. We need to know the *maximum* rate at which data will be lost. The models assume things, such as uncorrelated errors and bug-free software, that all experimental studies show are false. The models exclude most of the threats to which stored data is subject. And in those cases where similar claims, such as those for disk reliability (Schroeder and Gibson 2007; Pinheiro, Weber, and Barroso 2007), have been tested they have been shown to be optimistic. It is not reasonable to assume that these factors are negligible, nor that they affect all systems equally; the models thus provide an estimate of the *minimum* data loss rate to be expected.

Even if we believed the models, the MTTDL number doesn't tell us how much data was lost in the average data loss event. Is petabyte system *A* with a MTTDL of 10^6 years better than a similar size system *B* with a MTTDL of 10^3 years? If the average data loss event in system *A* loses the entire petabyte, where the average data loss event in system *B* loses a kilobyte, it would be easy to argue that system *B* was 10^9 times better.

It is clear that we need a better way to define and measure bit preservation performance. Mean time to data loss is not a useful characterization of how well a system stores bits through time.

Theory

In order to claim that “bit preservation is a solved problem” we would need three things we currently don't have:

- A specific requirement as to how well bits need to be preserved.
- A technique for measuring whether actual systems achieve the required level of bit preservation.
- Measurements of an actual system using the technique that confirm it meets or exceeds the requirement.

In this section we suggest a metric that would be more useful than MTTDL, and ask whether it is possible to characterize actual systems in terms of this metric.

Defining a Solution

The most abstract model of a bit preservation system is as a black box, into which a string of bits $S(0)$ is placed at time $T(0)$ and from which at subsequent times $T(i)$ a string of bits $S(i)$ can be extracted. The system is successful if $S(i) = S(0)$ for all i .

No real-world system can be perfect and eternal, so real systems will fail. The simplest model of these failures is analogous to the decay of radioactive atoms. Each bit in the

string independently is subject to a random process that has a constant small probability per unit time of causing its value to flip. The time after which there is a 50% probability that a bit will have flipped is the “bit half-life”.

The requirement of a 50% chance that a petabyte will survive for a century translates into a bit half-life of 8×10^{17} years. The current estimate of the age of the universe U is 1.4×10^{10} years, so this is a bit half-life approximately $6 \times 10^7 U$.

Measuring a Solution

Because current storage systems are extraordinarily reliable, measuring their bit half life involves observing very large numbers of bits for a very long time. If you wanted to take a year to measure whether a system met the petabyte-for-a-century requirement you might watch a thousand such systems, an exabyte of data. If the system were just good enough, you would see a single bit flip in just five of the systems.

Even if one were able to afford this experiment, doing so would be challenging. Data must be read from the system and compared with its expected value. Even if each bit is checked only once at the end of the year, the comparisons have to be performed with less than 1 chance in 10^{19} of any error.

In practice, estimates of bit half-life would have to be based upon the same models as estimates of MTTDL, and would thus share many of the same difficulties.

Assessment

There is no escape from the problem that the size of the data collections to be preserved and the times for which they must be preserved mean that experimental confirmation that the technology chosen is up to the job is not economically feasible. Even if it was the results would not be available soon enough to be useful. What this argument demonstrates is that, far from bit preservation being a solved problem, it is in a very specific sense an *unsolvable* problem. Even if we believed a system we developed was reliable enough, there are no feasible experiments that could confirm our belief in time to be useful.

Bit half-life is a more informative metric than MTTDL, because it is a measure of the reliability of the *data*, not a measure of the reliability of the *system* storing it. The data’s survival is what we care about. It thus captures the fact that the impact of a data loss event depends not just on when it happens, but also on how much data is lost. It is still far from ideal:

- Bits in real storage systems do not fail independently; they exhibit significant correlations in space and time (Bairavasundaram et al. 2008). These correlations make failure more likely than it otherwise would be. This observation doesn’t invalidate the simple “radioactive decay” model; it merely makes adequate bit half-life a necessary but not sufficient condition for a system to meet the requirement.
- Like MTTDL, it is a statistical estimate and thus, like MTTDL, it is not useful without an uncertainty interval.

- Because storage systems are so reliable, it is just as difficult to measure bit half-life as it is to measure MTTDL.

Practice

As enterprises such as Google (Chang et al. 2006) and institutions such the Sloan Digital Sky Survey (SDSS 2008) and the Large Hadron Collider (CERN 2008) collect petabytes of data with long-term value that must remain on-line to be useful, and as the annual cost of keeping a petabyte on-line is more than a million dollars (Moore et al. 2007), questions of the economics and reliability of storage systems have become the focus of researchers’ attention.

Storage Failures

Papers at the 2007 FAST conference used data from NetApp (Schroeder and Gibson 2007) and Google (Pinheiro, Weber, and Barroso 2007) to study disk replacement rates in large storage farms. They showed that the manufacturer’s MTTF numbers were optimistic. Subsequent analysis of the NetApp data (Jiang et al. 2008) showed that all other components contributed to the storage system failures, and:

“Interestingly, [the earlier studies] found disks are replaced much more frequently (2–4 times) than vendor-specified [replacement rates]. But as this study indicates, there are other storage subsystem failures besides disk failures that are treated as disk faults and lead to unnecessary disk replacements.”

Two studies, one at CERN (Kelemen 2007) and one using data from NetApp (Bairavasundaram et al. 2008), greatly improved on earlier work using data from the Internet Archive (Baker et al. 2006; Schwarz et al. 2006). They studied *silent data corruption* in state-of-the-art storage systems; events in which the content of a file in storage changes with no explanation or recorded errors.

The NetApp study looked at the incidence of silent storage corruption in individual disks in RAID arrays. The data was collected over 41 months from NetApp’s filers in the field, covering over 1.5×10^6 drives. They found over 4×10^5 silent corruption incidents. More than 3×10^4 of them were not detected until RAID restoration and could thus have caused data loss despite the replication and auditing provided by NetApp’s row-diagonal parity RAID (Corbett et al. 2004).

The CERN study used a program that wrote large files into CERN’s various data stores, which represent a broad range of state-of-the-art enterprise storage systems (mostly RAID arrays), and checked them over a period of 6 months. A total of about 9.7×10^{16} bytes was written and about 1.92×10^8 bytes was found to have suffered silent corruption, of which about 2/3 was persistent; re-reading did not return good data. In other words, about 1.2×10^{-9} of the data written to CERN’s storage was permanently corrupted within six months. We can place an upper bound on the bit half-life in this sample of current storage systems by assuming that the data was written instantly at the start of the 6 months and checked instantly at the end; the result is 2×10^8 or about $10^{-2}U$. Thus to reach the petabyte for a century

requirement we would need to improve the performance of current enterprise storage systems by a factor of at least 10^9 .

Surviving Storage Failures

Despite the manufacturer's claims, current research shows that state-of-the-art storage systems fall so many orders of magnitude below our bit preservation requirements that we cannot expect even dramatic improvements in technology to fill the gap. Maintaining a single replica in a single storage system is not an adequate solution to the bit preservation problem.

Practical digital preservation systems must therefore:

- Maintain more than one copy by *replicating* their data on multiple, ideally different, storage systems.
- Audit or (*scrub*) the replicas to detect damage, and repair it by overwriting the known-bad copy with data from another.

The more replicas and the more frequently they are audited and repaired the longer the bit half-life we can expect. This is, after all, the basis for the backups and checksums technique in common use. In fact, current storage systems already use versions of these techniques, for example in the form of RAID (Patterson, Gibson, and Katz 1988). Despite this the bit half-life they deliver is inadequate. Unfortunately adding the necessary inter-storage-system replication and scrubbing is expensive.

2007 cost figures from the San Diego Supercomputer Center (Moore et al. 2007) show that maintaining a single on-line copy of a petabyte for a year then cost about $\$1.5 \times 10^6$. A single near-line copy on tape cost about $\$5 \times 10^5$ a year⁵. These costs decrease with time, albeit not as fast as raw disk costs. The British Library estimates a 30% per annum decrease. Assuming that this rate continues for at least a decade, if you can afford about 3.3 times the first year's cost to store an extra replica for a decade, you can afford to store it indefinitely. So, adding a second replica of a petabyte on disk would cost about $\$3.5 \times 10^6$ and on tape would cost about $\$1.4 \times 10^6$. Adding cost to a preservation effort to increase reliability in this way is a two-edged sword; doing so necessarily increases the risk that preservation will fail for economic reasons.

Further, without detailed understanding of the rates at which different mechanisms cause loss and damage, it isn't possible to derive from a desired bit half-life the appropriate number of replicas⁶ and thus the cost implication of replication. At small scales the response to this uncertainty is to add more replicas, but as the scale increases this rapidly becomes unaffordable.

⁵SDSC reports that the 2008 costs are $\$1.05 \times 10^6$ and $\$4.2 \times 10^5$

⁶The number can be quite large; a study of paper journals (Yano 2008) found between 3 and 31 copies were needed to achieve loss probabilities over a century of between 10^{-3} and 10^{-6} given various plausible loss rates of the individual copies. The lower repairability of paper copies inflates these numbers, while their greater durability deflates them, as against digital copies.

Replicating among identical systems is much less effective than replicating among diverse systems. Identical systems are subject to common mode failures, for example caused by a software bug in all the systems damaging the same data in each. On the other hand, purchasing and operating a number of identical systems will be considerably cheaper than operating a set of diverse systems.

Each replica is vulnerable to loss and damage. Unless they are regularly audited they contribute little to increasing bit half-life. The bandwidth and processing capacity needed to scrub the data are both costly, and adding these costs increases the risk of failure. Custom hardware (Michail et al. 2005) could compute the SHA-1 (Nat 1995) checksum of a petabyte of data in a month, but doing so requires impressive bandwidth - the equivalent of three gigabit Ethernet interfaces running at full speed the entire month. User access to data in preservation systems is typically infrequent; they are therefore rarely architected to provide such high-bandwidth read access. System cost increases rapidly with I/O bandwidth, and the additional accesses to the data (whether on disk or on tape) needed for scrubbing themselves potentially increase the risk of failure.

The point of writing software that reads and verifies stored data in this way is to detect damage and exploit replication to repair it, thereby increasing bit half-life. How well can we do this? RAID is an example of a software technique of this type applied to disks. In practice, the CERN study (Kelemen 2007) looking at real RAID systems from the outside showed a significant rate of silent data corruption, and the NetApp study (Bairavasundaram et al. 2008) looking at them from the inside showed a significant rate of silent disk errors that would lead to silent data corruption. A study (Krioukov et al. 2008) of the full range of current algorithms used to implement RAID found flaws leading to potential data loss in all of them. Both this study, and another from IBM (Hafner et al. 2008), propose improvements to these algorithms but neither claim that they can eliminate silent corruption, or even accurately predict its incidence:

“while we attempt to use as realistic probability numbers as possible, the goal is not to provide precise data loss probabilities, but to illustrate the advantage of using a model checker, and discuss potential trade-offs between different protection schemes.” (Krioukov et al. 2008)

Thus although replication and scrubbing are capable of decreasing the incidence of data loss in current storage systems, they cannot eliminate it completely. And the replication and scrubbing software itself will contain bugs that can cause data loss. It must be doubtful that we can implement these techniques well enough to increase the bit half-life of systems with an affordable number of replicas by 10^9 .

It takes experiments with petabytes of storage to characterize the performance of current systems accurately. Even if we believed we had implemented replication and audit well enough to improve performance by 10^9 , we could not afford to do the experiments that would be needed to confirm it.

Policy

If bit preservation were a solved problem then it would be reasonable to expect that no bits would be lost. This is not the case; just as in paper archives preserved content in digital archives will be lost or damaged. Setting unreasonable expectations for the performance of our preservation systems, for example by continually making unsupported claims to have solved the bit preservation problem, is simply setting ourselves up to be perceived as failures.

If preserved bits will be lost, the question becomes how to invest the limited funds available to reduce the rate of loss as much as possible. It is a commonplace that if you can measure something you can improve it. The history of technology markets such as CPUs and graphics chips show that competition between vendors based on widely accepted standard benchmarks can drive rapid improvements in component cost-performance. Alas, although raw storage cost is easily measured and is the subject of effective competition to decrease cost per byte (Christensen 1997), long-term storage reliability is very hard to measure and the accepted metric for it is not very informative. Competition to reduce the cost of a given level of bit preservation is therefore much less effective.

It is in the interest of the digital preservation community to improve competition in their market. How could this be done?

- Agreement on a metric for bit preservation performance is an essential first step. It would be extremely valuable if it were possible to define one that was easily measurable, but this seems rather unlikely.
- Given this, it seems likely that numbers for bit preservation performance will continue to be generated by models. Achieving consensus on modeling techniques is important, especially as it appears that traditional techniques are running into difficulties (Storer et al. 2008; Elerath and Pecht 2007).
- These models will need agreed data. Better and more widely available data about the real world performance storage components is thus important. Realistic studies have only begun to be published, and they aren't yet based on shared metrics. The effort by Usenix and Carnegie-Mellon (Usenix 2008) to establish a repository for suitably anonymized data of this kind is to be commended.
- Storage systems are currently designed using completely inadequate models of how components fail. One problem is that these failures are highly correlated, making the models complex and difficult. A shared model of the threats against which bits need to be preserved, models of these threats, and data regarding their incidence is also important.
- Anecdotal evidence suggests that operator error and insider abuse are major causes of data loss in large storage farms; they are difficult to model or characterize. This is in part because sites are very reluctant to admit to data loss incidents. An anonymous incident reporting system modelled on NASA's Aviation Safety Reporting

System (NASA 2008) would be very valuable in understanding the mechanisms of, and defending against, these failures.

The fact that it is possible for digital information to be copied perfectly does not mean that it always will be. While perfection is not within the grasp of real-world engineers, improvement is always possible. However, improvement takes money, and without the research outlined above we are unable to make rational tradeoffs between the cost of preserving content to a given level of reliability and the cost of the losses implied by the given level.

Conclusions

As we have seen, the case that bit preservation is a solved problem rests on the conviction that the conventional techniques of backups and checksums are more than adequate to the scale of the problem. This conviction is odd. Press accounts (e.g. (Brodkin 2008)) of companies, presumably using the conventional techniques, nevertheless losing essential data are common. Awareness that systems frequently encounter scaling problems is also widespread, as is the expectation that the future demands for preserving digital content will be enormous.

But the case for bit preservation not being solved does not rest on this cognitive dissonance. It rests rather on the many orders of magnitude mismatch between the reliability requirements implied by society's expectations of the amount of data to be preserved and the length of time for which it should be preserved, and the observed performance of current storage hardware and software.

Were every bit to come adequately endowed with capital to provide guaranteed funds through time its preservation would not be a major concern, although it would still not be a solved problem. Like almost all engineering problems, bit preservation is fundamentally a question of budgets. Society's ever-increasing demands for vast amounts of data to be kept for the future are not matched by suitably lavish funds. Thus, absent a technological miracle, bit preservation is a problem with which we are doomed to struggle indefinitely.

Acknowledgements

Thanks are due to Michael Bax and the LOCKSS engineering team for critical readings of drafts of this paper, and to the staff of the San Diego Supercomputer Center for the discussions that started me thinking along these lines.

References

- Adams, D. 1978. *The Hitch-Hiker's Guide to the Galaxy*. British Broadcasting Corp.
- Bairavasundaram, L.; Goodson, G.; Schroeder, B.; Arpaci-Dusseau, A. C.; and Arpaci-Dusseau, R. H. 2008. An Analysis of Data Corruption in the Storage Stack. In *Proceedings of 6th USENIX Conf. on File and Storage Technologies*.
- Baker, M.; Shah, M.; Rosenthal, D. S. H.; Roussopoulos, M.; Maniatis, P.; Giuli, T.; and Bungale, P. 2006. A Fresh

- Look at the Reliability of Long-term Digital Storage. In *Proceedings of EuroSys2006*.
- Brodin, J. 2008. Loss of customer data spurs closure of online storage service 'The Linkup'. *Network World*. 11th Aug.
- CERN. 2008. Worldwide LHC Computing Grid. <http://lcg.web.cern.ch/LCG/>.
- Chang, F.; Dean, J.; Ghemawat, S.; Hsieh, W. C.; Wallach, D. A.; Burrows, M.; Chandra, T.; Fikes, A.; and Grube, R. E. 2006. Bigtable: A Distributed Storage System for Structured Data. In *Proceedings of the 7th Usenix Symp. on Operating System Design and Implementation*, 205–218.
- Christensen, C. M. 1997. *The Innovator's Dilemma: When New Technologies Cause Great Firms to Fail*. Harvard Business School Press.
- Corbett, P.; English, B.; Goel, A.; Grcanac, T.; Kleiman, S.; Leong, J.; and Sankar, S. 2004. Row-Diagonal Parity for Double Disk Failure Correction. In *3rd Usenix Conference on File and Storage Technologies*.
- Elerath, J. G., and Pecht, M. 2007. Enhanced reliability modeling of raid storage systems. In *DSN '07: Proceedings of the 37th Annual IEEE/IFIP International Conference on Dependable Systems and Networks*, 175–184. Washington, DC, USA: IEEE Computer Society.
- Engler, D. 2007. A System's Hackers Crash Course: Techniques that Find Lots of Bugs in Real (Storage) System Code. In *Proceedings of 5th USENIX Conf. on File and Storage Technologies*.
- Hafner, J. L.; Deenadhayalan, V.; Belluomini, W.; and Rao, K. 2008. Undetected disk errors in RAID arrays. *IBM J. Research & Development* 52(4/5).
- Jiang, W.; Hu, C.; Zhou, Y.; and Kanevsky, A. 2008. Are Disks the Dominant Contributor for Storage Failures? A Comprehensive Study of Storage Subsystem Failure Characteristics. In *Proceedings of 6th USENIX Conf. on File and Storage Technologies*.
- Kelemen, P. 2007. Silent Corruptions. In *8th Annual Workshop on Linux Clusters for Super Computing*.
- Krioukov, A.; Bairavasundaram, L. N.; Goodson, G. R.; Srinivasan, K.; Thelen, R.; Arpaci-Dusseau, A. C.; and Arpaci-Dusseau, R. H. 2008. Parity Lost and Parity Regained. In *Proceedings of 6th USENIX Conf. on File and Storage Technologies*.
- Michail, H. E.; Kakarountas, A. P.; Theodoridis, G.; and Goutis, C. E. 2005. A low-power and high-throughput implementation of the SHA-1 hash function. In *Proceedings of the 9th WSEAS International Conference on Computers*.
- Moore, R. L.; D'Aoust, J.; McDonald, R. H.; and Minor, D. 2007. Disk and Tape Storage Cost Models. In *Archiving 2007*.
- NASA. 2008. Aviation Safety Reporting System. <http://asrs.arc.nasa.gov/>.
- National Institute of Standards and Technology (NIST), Washington, D.C., USA. 1995. *Federal Information Processing Standard Publication 180-1: Secure Hash Standard (SHA-1)*.
- Parkinson, C. N. 1957. *Parkinson's Law*. Buccaneer Books.
- Patterson, D. A.; Gibson, G.; and Katz, R. H. 1988. A Case for Redundant Arrays of Inexpensive Disks (RAID). In *Proceedings of the ACM SIGMOD International Conference on Management of Data*, 109–116.
- Pinheiro, E.; Weber, W.-D.; and Barroso, L. A. 2007. Failure Trends in a Large Disk Drive Population. In *Proceedings of 5th USENIX Conf. on File and Storage Technologies*.
- Popper, K. 1959. *Logic of Scientific Discovery*. Hutchinson. Footnote *5, Chapter X.
- Prabhakaran, V.; Agrawal, N.; Bairavasundaram, L.; Gunawi, H.; Arpaci-Dusseau, A. C.; and Arpaci-Dusseau, R. H. 2005. IRON File Systems. In *Proceedings of the 20th Symposium on Operating Systems Principles*.
- Rosenthal, D. S. H.; Robertson, T. S.; Lipkis, T.; Reich, V.; and Morabito, S. 2005. Requirements for digital preservation systems: A bottom-up approach. *D-Lib Magazine* 11(11).
- Schroeder, B., and Gibson, G. 2007. Disk failures in the real world: What Does an MTTF of 1,000,000 Hours Mean to You? In *Proceedings of 5th USENIX Conf. on File and Storage Technologies*.
- Schwarz, T.; Baker, M.; Bassi, S.; Baumgart, B.; Flagg, W.; van Imngen, C.; Joste, K.; Manasse, M.; and Shah, M. 2006. Disk Failure Investigations at the Internet Archive. In *Work-in-Progress Session, NASA/IEEE Conf. on Mass Storage Systems and Technologies*.
- SDSS. 2008. The Sloan Digital Sky Survey. <http://www.sdss.org/>.
- Storer, M. W.; Greenan, K. M.; Miller, E. L.; and Voruganti, K. 2008. Pergamum: Replacing Tape with Energy Efficient, Reliable, Disk-Based Archival Storage. In *Proceedings of 6th USENIX Conf. on File and Storage Technologies*.
- Sun Microsystems. 2006. Sales Terms and Conditions, Section 11.2. http://store.sun.com/CMTemplate/docs/legal_terms/TnC.jsp#11.
- Sun Microsystems. 2008. ST5800 presentation. Sun PASIG Meeting.
- Talagala, N. 1999. *Characterizing Large Storage Systems: Error Behavior and Performance Benchmarks*. Ph.D. Dissertation, CS Div., Univ. of California at Berkeley, Berkeley, CA, USA.
- Usenix. 2008. The computer failure data repository (CFDR). <http://cfdr.usenix.org/>.
- Williams, P.; Rosenthal, D. S. H.; Roussopoulos, M.; and Georgis, S. 2008. Predicting the Archival Life of Removable Hard Disk Drives. In *Archiving 2008*.
- WWPDB. 2008. Worldwide Protein Data Bank. <http://www.wwpdb.org/>.
- Yano, C. 2008. How Many Journal Copies? A Preliminary Report. Presentation to ALA.

The Modeling System Reliability For Digital Preservation: Model Modification and Four-Copy Model Study

Yan Han, Chi Pak Chan

The University of Arizona Libraries
1510 E University Blvd, Tucson, AZ, USA, 85721
{yhan, cpchan}@email.arizona.edu

Abstract

Research has been studied to evaluate the reliability of storage media and the reliability of a computer backup system. In this paper, we use the Continuous Time Markov Chain to model and analyze the reliability of a computer backup system. We propose a modified model from that of the Constantopoulos, Doerr and Petraki [1]. We analyze the difference, show computational results, and propose new input parameters (e.g. time to repair) for the model from our experience. Further we developed a four-copy data model to test if it fulfills the sample reliability rate set by the RLG-NARA. The modeling process can be applied to construct models for computer preservation systems using different storage media. The reliability of constructed models can be calculated so that preservation institutions can have quantitative data to decide their preservation strategies.

1. Introduction

Traditional preservation techniques have focused on longevity of the media since the only requirement has usually been human readability. With a growing number of born-digital data and digitized materials there is an urgent need for research on digital preservation. Unlike traditional preservation strategies, digital preservation fundamentally changes the nature and process of preservation while considering issues related to media, storage, access, representation, and authentication. Digital preservation is more complex, not only because of the information encoded in various IT standards or protocols, but also related to its context: metadata management and higher level of policy issues.

Digital preservation has two related components: physical and logical preservation. Physical preservation for digital assets is similar to preserving analog materials and ensuring bit-streams to be readable from storage media. Logical preservation is more complex because it requires technology and processes to ensure that bit-streams are renderable and accessible for computers and humans. This

paper discusses using Continuous Time Markov Chain to measure capacity of physical preservation, including modeling, analyses, and comparisons between the CDP's model [1] and our modified model. We suggest new input parameters such as time to repair for the model and construct a four-copy backup system.

2. Related Research

Digital files are vulnerable to corruption due to multiple reasons such as failed storage media, outdated backup, obsolete recording/reading devices, neglected human errors, and undesirable disasters. The longevity of digital storage media has been a subject of interest to librarians and archivists. In 2002, National Archives and Records Administration (NARA) directed a study of high density magnetic tapes life expectancy and revealed tapes can have a life expectancy of 50 -100 years [8][9]. The Library of Congress completed an unpublished report to study prerecorded compact discs (CD-ROMs). Both the National Institute of Standards and Technology (NIST) in 2004 [6] and Canadian Conservation Institute in 2005 published reports of life expectancies of recordable CDs (CD-Rs), rewriteable CDs (CD-RWs), and recordable DVD (DVD-Rs). All the studies show that higher deterioration for optical and magnetic media, when exposure to high temperature and humidity condition.

To establish a process to ensure long-term sustainability for digital collections, Research Library Group (RLG) and United States National Archives and Records Administration (NARA) released a report for evaluating a trusted digital repository. The report covers critical digital preservation issues, including physical and logical preservation for long term preservation. The report states that "D1.5 Repository has effective mechanisms to detect data corruption or loss" [3] and illustrates a sample reliability rate: "if the policy were the repository could not lose more than 0.001% of the collection per year..." [3] The quantitative data allows preservation institutions and certificate issuing organizations to measure the capacity of a trusted digital repository. Since 1999, Lots of Copies Keep Stuff Safe (LOCKSS) [7] advances digital preservation research and receives tremendous success in libraries and publishers. LOCKSS

is a peer-to-peer open source software to convert a PC into a digital preservation node, creating low cost, persistent, and accessible copies of web-based data. Since LOCKSS is a peer-to-peer system, it is an innovation to just show the concept of “the more, the better”. However, LOCKSS might not be appropriate for close data.

In 2005, Constantopoulos, Doerr and Petraki (CDP) published a paper [1] to introduce a reliability model that uses the Continuous Time Markov Chain to measure the reliability of a computer preservation system.

3. Methodology

As more and more preservation institutions are involved with digitization, and at the same time anticipating growing needs of preserving born-digital materials, it is critical to have quantitative study on the reliability of a computer backup system so that preservation institutions can base on outputs from quantitative analysis to make decisions for long-term preservation. In the CDP’s paper [1], it was calculated that a typical computer backup system with three-copy of data (two disks and one tape) has a reliability rate of 67.46% in 1000 years. Since this paper does not provide the unreliability rate of one year, we drew the system and calculated that the system’s unreliability rate is 0.033%. This result obviously does not meet the 0.001% unreliability rate illustrated by the RLG-NARA report. Is the reliability modeling appropriate? If we develop a four-copy data model, will it fulfill the RLG-NARA’s required 0.001% unreliability rate? Is it possible that the modeling can be easily extended to more copies of data and different storage media?

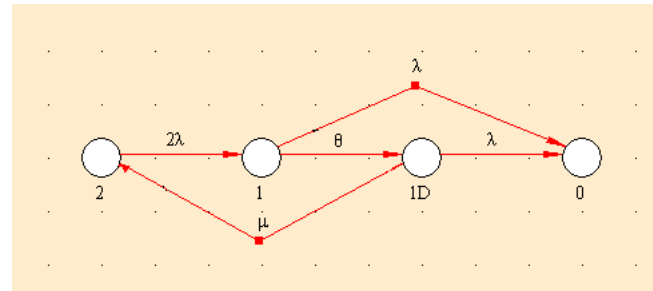
Continuous Time Markov Chain (CTMC) is used to analyze lifetime and reliability rate of a backup system. Computer system components such as disk, tape, and other forms of storage media could break down at any time due to depreciation of the components or some unexpected external factors such as earthquake or flooding, which can take place at a random time. On the other hand, the recovery process of a component is approximately a continuous process. Moreover, it is reasonable that the probability of a system’s next state bases only on current state of the system. Therefore, CTMC is an appropriate methodology to analyze system continuous failure/recovery processes and state status of the whole system. Inspired by the RLG-NARA’s report and the CDP’s paper, we conduct further research on this topic.

3.1 Markov Modeling The preservation policy is: for each digital file we create one or more copies in disk, tape, or other forms of storage media, and if detecting a failure of disk, tape, or other forms of storage media, we replace

them. We assume that the preservation policy is consistent over the time.

In this case, we analyzed a mirrored computer system with two copies of data in hard disks. We are interested in finding out the reliability of this system. The CDP’s paper has already described the process of constructing the Markov chain for the two-copy system [1]. We had the same result. (See Figure 1)

Figure 1: Modeling a two-copy system



Where

- State 2: Both disks function properly.
- State 1: One of the disks has failed, but has not detected yet.
- State 1D: One of the disks has failed and the failure has been detected.
- State 0: Both disks failed (absorbing state). Therefore, the data is not recoverable.

Initially the system starts at state 2 (2 copies function properly) and each disk has a failure rate; assuming that both have the same failure rate λ , the rate for the system going from State 2 to State 1 is 2λ , as shown on arc (2,1). There is a rate regarding the detection of the failure disk, which is θ shown on arc (1,1D). Moreover, there is a possibility that the other functioning disk fails even before failure of the failed disk has been detected, and the rate for the system going from State 1 to State 0 is λ as shown on arc (1, 0), which results in the failure of the whole system and the data never being recovered. Similarly, at State 1D, the failure of the disk has been detected and is repaired. There is a possibility that the system can fail (i.e. from State 1D to State 0) and the rate is λ . There is a possibility the system recovers to its initial state (2 disks) by recovering the failed disk as shown on arc (1D, 2) with rate μ .

3.2 Our Experience about Input Parameters and Storage Media.

The CDP’s paper [1] conducted experiments to study the above parameters such as mean time to failure (MTTFdisk), mean time to repair (MTTRdisk), and mean time to detect failure (MTTFDdisk) for their modeling. The University of Arizona Libraries had a few server disk failures and tape failures in the past. Our experience shows that it takes us about 25 hours to restore 10TB data back to a storage (hard disks) server, if the backup policy requires systems administrators to recover the data as soon as possible. This process includes reinstalling Operating System (OS) and

Copyright © 2008, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

copying 10TB of data using 1000 Mbps network connection. It is true that less data takes less time to repair. Therefore, MTTRdisk depends on the amount of data and computer backup policies. For MTTFDdisk we can detect hard disk failure right away, because modern OS automatically sends emails/text messages to us and server vendor when detecting failed hard disks. Server vendors such as Dell offer 24x7 replacement service plan to deliver new hard disks to us, and we use Dell's 4-hour replacement plan. If the storage server is critical, we can upgrade our service plan to get quicker service. Our MTTRdisk is 25 hours and MTTFDdisk is 4 hours, compared to the CDP's 50 hours of MTTRdisk and 14 days of MTTFDdisk. Our experience on MTTFDtape is different from MTTFDdisk. Currently we do not have a tape library and thus our systems administrators have to manually change tapes. This slows down time to detect and repair tapes. Using restoring 10TB of data as an example, our MTTRtape is 60 hours, and MTTFDtape is about 60 days. In addition, the costs and benefits of storage media and staffing should not be ignored. In practice, tapes require more staffing time to handle, more time to access data, and is less reliable, but they are easy to store offsite and cheap in terms of cost per GB. Compared to hard disks and magnetic tapes, CDs and DVDs are limited in storage size, require frequent human handling when reading data, and are usually not rewriteable. Due to the above disadvantages, in 2004 we made a decision to remove optical media for permanent storage at the University of Arizona Libraries.

The input parameters from the CDP's model are close to what we measured from real life experience except MTTFDdisk. To best comparing the differences between the CDP's model and our modified model, we use the same input parameters.

$$\begin{aligned}
 MTTF_{disk} &= 1/\lambda = 3 \text{ years} \Rightarrow \lambda = 1/3 \text{ per year} \\
 MTTR_{disk} &= 1/\mu = 50 \text{ hours} \Rightarrow \mu = 175.2 \text{ per year} \\
 MTTFD_{disk} &= 1/\theta = 14 \text{ days} \Rightarrow \theta = 365/14 \text{ per year}
 \end{aligned}$$

Let

$m_i = E[\text{time before absorption} \mid \text{the system starts from state } i]$
Based on the CTMC model, we have the following set of equations:

$$\begin{aligned}
 m_2 &= 1/2\lambda + m_1 \\
 m_1 &= [\lambda/(\theta + \lambda)] \times 1/\lambda + [\theta/(\theta + \lambda)] \times m_{1D} \\
 m_{1D} &= [\lambda/(\mu + \lambda)] \times 1/\lambda + [\mu/(\mu + \lambda)] \times m_2
 \end{aligned}$$

Which are reduced to the following:

$$\begin{aligned}
 m_2 &= 1/2\lambda + 1 \times m_1 \\
 m_1 &= 1/(\theta + \lambda) + \theta/(\theta + \lambda) \times m_{1D} \\
 m_{1D} &= 1/(\mu + \lambda) + \mu/(\mu + \lambda) \times m_2
 \end{aligned}$$

Solve these and we get:

$$m_2 = \frac{(\theta + \lambda)(\mu + \lambda) + 2\lambda(\mu + \lambda) + 2\lambda\theta}{2\lambda^2(\theta + \mu + \lambda)}$$

Petraki's paper has $m_2 = \frac{(\theta + \lambda)(\mu + \lambda) + 2(\mu + \lambda) + 2\lambda\theta}{2\lambda^2(\theta + \mu + \lambda)}$ [2], which might

be a typographical error. The expected TTF_{system} m_2 is 106.46 years, which means that the two-copy system is expected to crash after 106.46 years. To verify the result, we used a software package called SHARPE to model this Markov Chain. The result is exactly the same as we got from the above formula: 106.46 years.

3.3 Modeling a three-copy model The CDP's paper [1] also described the process of extending the system by adding another backup copy. Their example is to add magnetic tapes for an additional copy of data. Other media such as CD and DVD can also be used with appropriate rates $(\lambda_3 \theta_3 \mu_3)$.

Our Markov model shown in Figure 3 on a three-copy system (2 in disk and 1 in tape) is similar to that of CDP's model [1] shown in Figure 2, but we propose some modifications which we think are more realistic in real life situations and less risky in preventing a backup system from ending at the absorbing state. In the figures, we use *(DiskCopiesFunctioningDiskCopiesFailureDetected_TapeCopiesFunctioningTapeCopiesDetected)* notation to represent state status. *DiskCopiesFunctioning* represents the number of copies of data functioning, while status *DiskCopiesFailureDetected* can be either *NULL* or *D* (meaning failure detected). Figure 2 shows CDP's model for three-copy system.

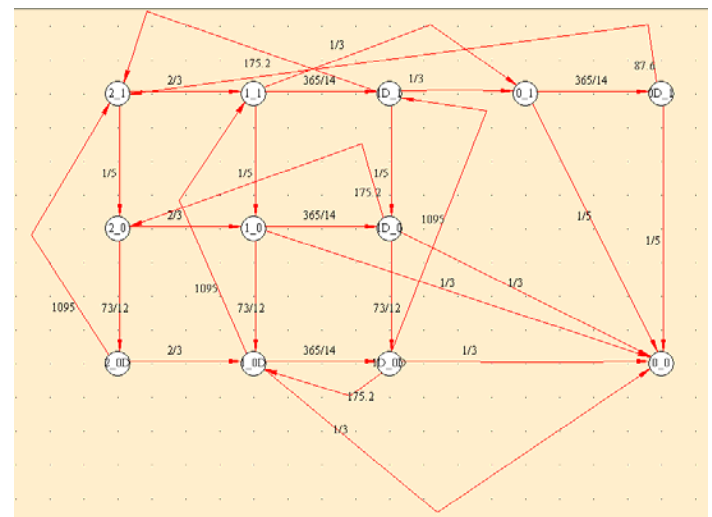


Figure 2: CDP's Markov model for a three-copy system (2 in disk and 1 in tape)

We believe that repairing failed copies one at a time is more realistic due to resource limitations. Therefore, we suggest that whenever there is/are failed copy/copies detected, only one failed copy will be repaired at a time to

$MTTF_{system}$ and reliability rate of our model are higher than that of CDP's, because our model allows repairing one copy of data at a time (i.e. arcs from 0D_1 to 1D_1 and 1D_0D to 2_0D). The unreliability rate in 1 year of our three-copy model is 0.0324%.

4. A CTMC model for a four-copy system

While the three-copy model does not fulfill the 0.001% unreliability yearly rate set forth by the RLG-NARA's report, we extend the model for four-copy of data (2 in disk and 2 in tape). The model is shown as follows (Figure 4). Again, as we have done for the three-copy model, the four-copy model allows repairing one failed copy at a time, but does not allow repairing multiple failed copies at a time. Using the same input parameters (e.g. $MTTF_{disk}$, $MTTF_{tape}$) as above, Computational output of the model for four-copy of data is as follows: $MTTF_{system}$ is about 4.238×10^4 years, reliability rate is 97.67% after 1000 years and the unreliability rate in 1 year is 0.001693% which nearly fulfills the RLG-NALA's requirement.

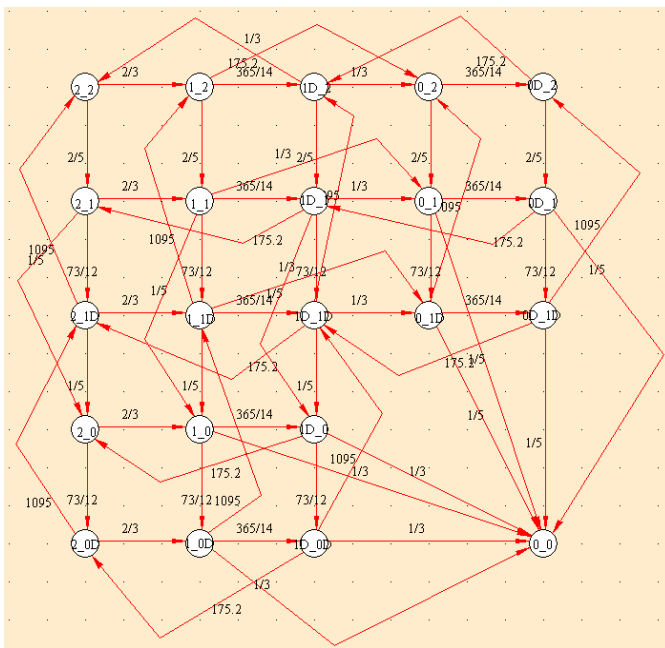


Figure 4: Our Markov modeling for a four-copy system (2 in disk and 2 in tape)

When feeding our input parameters (e.g. $MTTF_{disk}$, $MTTFD_{disk}$), the four-copy system fulfills the RLG-NALA's requirement. This makes sense because our $MTTFD_{disk}$ takes much less time to detect failures and our $MTTF_{disk}$ is 50% quicker to repair failed disks. One can also construct different four-copy systems such as 3-disk-1-tape and 4-disk. This of course proves the concept of

“the more, the better”, but the model gives a way to demonstrate how much better. Input parameters ($MTTF$, $MTTR$, and $MTTFD$) are critical to the reliability of a backup system. How much effort does each parameter play can be a following topic for research. The result can help an institution to tune its preservation policy.

5. Discussions

Inspired by the RLG-NARA's report and the CDP's paper [1], we have developed a modified CTMC model, which we think is more realistic in practice. We took a close look at the CDP's model and believe that the CDP's model is sound except handling repairing failed copies. We believe that repairing one copy at a time is more realistic in a real life situation.

Our experience shows that our $MTTF$, $MTTR$, and $MTTFD$ in disk are different. We researched optical storage media and discuss pitfalls of CDs and DVDs, and recommend not to use them for permanent storage. Tapes also have limitations when considering dropping cost and growing capacity of disks. We are considering reducing using tapes for backup.

Based on the rationale to build the CTMC model for the three-copy backup system, we've also developed a model for a four-copy backup system to test whether it can fulfill the sample reliability rate set by the RLG-NARA paper. With CTMC technique, reliability of a computer preservation systems can be calculated so that preservation institutions can use quantitative basis to decide their preservation strategies (e.g. how many copies of data are needed, forms of storage media, preservation policies) to ensure readability of bit-streams.

6. Acknowledgements

We would like to thank Professor Kishor S. Trivedi of Department of Electrical and Computer Engineering at Duke University for offering software SHARPE. We also would like to thank Qu Miao (who was a Computer Science graduate student at the University of Arizona) for providing interesting derivation of m_2 (expected time to failure of the system) using conditional expectations.

References

- [1] Constantopoulos, P., Doerr, M., and Petraki, M. 2005. *Reliability modelling for long term digital preservation*. <http://delos-wp5.ukoln.ac.uk/forums/dig-rep-workshop/constantopoulos-1.pdf>
- [2] Petraki, M. 2005. *Evaluating the reliability of system configurations for long term digital preservation*, Master of Science Thesis, Dept. of Computer Science, University of Crete. <http://www.ics.forth.gr/isl/publications/paperlink/Petraki.pdf>
- [3] RLG and NARA, 2005. *Audit Checklist for Certifying Digital Repositories*. <http://digitalarchive.oclc.org/da/ViewObjectMain.jsp?fileid=0000059075:000003279368&reqid=610>
- [4] Ross, Sheldon M. 2003. *Introduction to probability models*, 8th ed. San Diego, CA: Academic.
- [5] Ross, Sheldon M. 1996. *Stochastic Processes*, 2nd ed. New York: Wiley.
- [6] Slattery, O., Lu, R., Zheng, J., Byers, F., and Tang, X. 2004. *Stability Comparison of Recordable Optical Discs – A Study of Error Rates in Hash Conditions*. Journal of Research of the National Institute of Standards and Technology, vol, 109. pp. 517-524.
- [7] Stanford University Libraries. LOCKSS. <http://www.lockss.org>
- [8] Judge, J.S., Shmidt, R.G., Weiss, R. D., and Miller, G. 2003. *Media Stability and Life Expectancies of Magnetic Tape for Use with IBM 3590 and Digital Linear Tape Systems*. Proceedings of the 20th IEEE/11th NASA Goddard Conference on Mass Storage Systems and Technologies (MSS'03). http://storageconference.org/2003/papers/15_Judge-Media.pdf
- [9] Weiss, R.D. 2002. *Environmental Stability Study and Life Expectancies of Magnetic Media for Use with IBM 3590 and Quantum Digital Linear Tape Systems*. <http://www.archives.gov/research/electronic-records/magnetic-media-study.pdf>

Ways to deal with complexity

Christian Keitel

Landesarchiv Baden-Württemberg
Staatsarchiv Ludwigsburg, Arsenalplatz 3
71638 Ludwigsburg, Germany
christian.keitel@la-bw.de

Abstract

Several ways to deal with complexity are discussed. An archive can handle the matter by keeping the number of the single elements in the core areas of digital preservation down. The numbers of action types during the ingest process, of metadata and journals could be reduced. A preservation model for analogue and digital records is outlined. By keeping complexity down, it's easier to see what digital and analogue archiving have in common. Instead of seeing two totally different worlds (here is the old one, there is the new one), one can shift to a less revolutionary view. This makes it possible to fall back on the considerable implicit knowledge of the existing memory institutions. From the perspective of the whole archive, there are strong arguments for reducing complexity and keeping digital and analogue things together whenever possible.

Complexity can also be handled by cooperation. The Landesarchiv Baden-Württemberg appreciates the opportunity to use the software tools DROID and JHOVE. The BOA project is a further example for a venture in website archiving that is maintained by libraries and archives collaboratively.

Cooperation and the reduction of complexity are the two most promising ways to enable small and medium sized archives to start with digital preservation. Automation seems to be a good thing whenever it can be achieved, but until this stage is reached, the single steps and the standards which must be followed often are extremely complex.

Complexity matters

Over the past years considerable progress has been achieved in the area of digital preservation. PREMIS explains which preservation metadata should be kept; METS describes how to build an information package; PAIMAS lists nearly 90 steps for the ingest process and DRAMBORA enumerates the possible risks of digital archiving on more than 200 pages. These standards or guidelines have resolved many of the open questions. On the basis of these results and foundations, it should be easy to build a digital archive. Therefore it is striking that these achievements have not been followed by a significant increase in the number of digital archives. Although many memory institutions have assumed the task of securing and preserving digital objects, only some of them are actually doing this. How can this discrepancy between the progress of digital archiving and the widespread failing of implementations be explained? Has there at least been a public discussion of this problem?

Three observations may contribute to the search for an answer:

1. Beyond doubt, the named standards and guidelines are all extremely helpful. Their detailed information addresses both general and special problems. But it is a hard job to extract from these texts some general hints how to start with digital archiving. This task is even harder for a beginner in digital archiving.
2. The communities of the traditional archivists on the one hand and the digital archivists on the other hand are deeply divided. Each community is oblivious of the other. Hence, the implicit knowledge of a still existing memory institution is only rarely taken into account when setting up standards for digital archiving.
3. Standards are usually devised by members of big institutions like national archives or national libraries. Once again it must be stated that these are very valuable contributions. But are they equally applicable to smaller archives or libraries?

For many memory organizations, complexity is one of the most serious impediments to start with digital preservation. The extensiveness of the standards and the large number of articles published raise the suspicion among librarians and archivists that digital preservation is something nobody can really cope with, nobody or only the biggest memory institutions. It seems to scare all people who are supposed to establish digital archives but so far haven't started. But complexity is more than just a psychological problem. In the long term, complexity makes preservation more expensive and less feasible. So it is worthwhile to think about how we can deal with it.

One possible answer to this question is cooperation. Cooperation takes centre stage in many articles, projects and conferences. Although the necessity for cooperation can't be overestimated, it is not the only way to deal with complexity. Archives can also try to reduce it. For example, many specialists in digital preservation keep the number of their archival formats down. Thus, they reduce the complexity of digital preservation. But beyond this example there is remarkably little discussion about this option to deal with complexity. A third possibility to reduce complexity would be automation.

Some of the recently published recommendations can be seen as a preparatory work for further automation. The preservation manager simply presses a button and all the complex work will be done by the machine. But defining such machines seems rather complicated, as you can see, for example, at MoReq2. As a result, the recommendations are growing more and more complex while the implementations (the machines) are still out of sight.

Summing up, complexity seems to be a serious obstacle on the path to digital archiving. This paper describes some of the ways in which the Landesarchiv Baden-Württemberg tries to deal with it. The results presented below were devised in the course of the project “Digital Archive in the Landesarchiv Baden-Württemberg”, running from the end of 2005 until 2009.

Standards on Ingest

The Open Archival Information System, better known as OAIS, describes six functional entities: Ingest, Data Management, Archival Storage, Preservation Planning and Access. Altogether, the standard describes about 30 functions. In the summary chart these are connected with each other by almost 70 (68) arrows. What does this mean for someone trying to set up a digital archive? Even if each arrow corresponds to only one task, there still is a lot of work to be done.

For the ingest area OAIS specifies the following functions:

- receiving SIPs
- performing quality assurance on SIPs
- generating an Archival Information Package (AIP)
- extracting Descriptive Information from the AIPs for inclusion in the archive database and
- coordinating updates to Archival Storage and Data Management.

The functions are characterised in a highly abstract way and they are not ordered chronologically.

Two years after the publication of OAIS, the Management Council of the Consultative Committee for Space Data Systems (CCSDS) released a second recommendation: The Producer-Archive Interface Methodology Abstract Standard. PAIMAS gives a more detailed view of the relationships and interactions between a producer and an archive. Although the specification covers only the first stage of ingest, it still needs 86 steps to describe the transfer of a record from the producer to the archive. This is divided into four phases:

- Preliminary Phase (46 steps)
- Formal Definition Phase (36 steps)
- Transfer Phase (2 steps) and
- Validation Phase (2 steps).

Speaking of “phases” implies a chronological order of the single steps. In fact, the recommendation starts with

the identification of the contact persons and the exchange of general information (P-1 and 2). PAIMAS here is much more concrete than OAIS, but can the recommendation be understood as a true construction plan for a digital archive? There are at least two arguments against this assumption: Firstly, it seems nearly impossible to go through 86 steps just to run the first half of the ingest process, i.e. to transfer an object to the digital archive. Secondly, the concept lacks flexibility. The strict chronological order of the single steps forces the readers to go gradually forward. As each step is based on another, their order can’t be changed. The catalogue of PAIMAS therefore needs further transformation to become a construction plan for the Ingest to a digital archive.

An interesting proposal was made last year by the members of the Australasian Digital Recordkeeping Initiative (ADRI). They designed a Submission Information Package. Deliberately, a number of questions are not addressed. Nothing is said about the high level transfer process or the low level protocols or the physical transfer mechanisms. In other words, ADRI has done two things: On the one hand, they concentrated on a decisive part of the ingest process, and on the other hand, the result of their work (the SIP) allows each institution a lot of flexibility.

Ingest

Following OAIS, many institutions are forced to preserve their digital information in a way which is similar to the work of the traditional archives. Therefore, if a traditional, paper-based archive goes digital and plans to preserve digital information, many of the functions mentioned in OAIS are already well known. A traditional archive can thus refer to the implicit knowledge of its staff. So, it is not important to recall all the steps of PAIMAS. Anyway, if one takes into account the implicit knowledge of a memory organisation, prescribing a fixed ingest process seems to be rather the wrong way. Every archivist could name cases, in which the normal sequence of steps can’t be maintained. On the other hand, it is obvious that the traditional ingest process is not sufficient for the transfer of digital objects. Hence, a fundamental analysis of the whole process was done during the project, aiming at maintaining both the flexibility for the archivists and the manageability of the ingest process. As a main result, a distinction between action types and process steps was introduced. An action type can be seen as a tool: You can use it whenever it is necessary in a single process step. One action type can be used in different process steps. Of course, there is a typical way to proceed in the ingest process, so a list of the normal sequence of the single process steps was drawn up. But it is important to point out that nobody is forced to follow them in the order listed.

How many action types should be distinguished? Within the context of a traditional archive their number can be reduced to four: Appraisal, inventory taking, transfer and validation. These are the action types which are essential for the Ingest of digital information.

information is taken automatically, e.g. who has done the inventory making or a validation, when it was done and what the results were. Some information can be added by the archivist, e.g. why a single step was taken or how an export was done up to the insertion of a SQL statement. According to OAIS the ingest process ends with the transfer of the objects to the area of Archival Storage. At this moment, IngestList contains a full journal of the entire process. Due to the MD5 file and the amount of related information, which allows many cross checks, the information contained in the journal as well as the single lists give good evidence about trustworthiness. At the same time, IngestList doesn't require a fixed sequence of single steps. Therefore, the archivists are as flexible as they are with the Ingest of paper records.

Archival objects

Traditional paper based archives only work with one kind of objects: A paper record comprises the logical information and the physical carrier in an inseparable way. Each object has its defined limits.

But things are different in the digital world. PREMIS has shown us the fundamental split between logical and physical objects. According to this standard, the latter fall into three subtypes. Digital preservation itself seems to be much more complicated than the preservation of analogue materials, and the two tasks seem to be completely separated. But on the other hand, with microfilming and digitising of analogue objects we have already crossed the boundaries of the analogue world. Therefore, some important questions that came up were whether all kinds of records could be unified in one system of description and whether this could be done in a fairly simple way.

PREMIS distinguishes between representation, file and bitstream. A representation embodies the logical information (intellectual entity) and can contain files and bitstreams, whereas a file can contain just one or more bitstreams. Hence, a bitstream must depend on a file, but a file can depend on a representation or immediately on the intellectual entity. So, one entity of the analogue world is opposed to four objects in the digital world.

First, let us look at the world of the digital objects. The representation allows us to name exactly that bundle of files which represents a record. For this reason it is obvious that in many cases we need the concept of representation. But the question was: Is it acceptable that some files depend immediately on the logical object and others are part of a representation? Making use of representations means preserving different versions of a digital object over time. Making no use of representations in the case of a migration means either overwriting the old file or renaming the new file. Is it possible to preserve millions of files over centuries, some of which with their predecessors preserved, others without them, and still others bound together within a representation? To adopt this model would increase the number of different preservation paths and therefore also the complexity of future decisions on preservation. So, in this case we argued against flexibility because we didn't want to allow totally different preservation paths. All our

digital records therefore have at least one representation. The digital representations consist of files, but a file can't depend immediately on an information object.

Our second question was: Is it possible or even recommendable to introduce the representation model for the analogue born objects as well? Obviously we live in a time of copies. If you want to preserve some of the copies of analogue materials for a long time and make them searchable you have to think about the representation model. For these reasons, the Landesarchiv Baden-Württemberg has decided that all records (digital and analogue) should have at least one representation. Both representation and intellectual entities are listed in our finding aids (the OAIS area Data Management), whereas the digital files aren't shown there. So, analogue and digital materials are described in the finding aids together and in the same way.

The representation model presented above allows us to start our preservation activities for all kinds of objects in the area of Data Management. Some of the analogue born objects have to be preserved (e.g. a parchment charter), others can't be preserved (e.g. a drawing in glassine). Many of them are listed alongside another representation. Seeing a logical object with more than one representation means seeing different opportunities for preservation. Thus, preservation planning can almost be seen as information management. It has to be stated, though, that the material properties of a medieval charter in this context are a part of the information as well. However, there is a common entry point for all archival objects in our system, and the number of preservation paths and the complexity of the preservation have been remarkably reduced.

Archiving system

Looking for a repository system can cause severe headaches. First you have to define your requirements. Which objects should be archived? There is a great diversity of digital object types which archives may want to preserve. Most of these objects are embedded in hierarchical structures, which are not standardised but quite flexible. See for example the classification schemes and their distinction between series, files, subfiles, records and documents as described by MoReq2. These structures should be preserved together with the records themselves, but it's not easy to define the exact borders of each object. As a result, we have to maintain different objects complete with the logical links between them.

Another requirement was to keep the Archival Information Package within the file system and to use a database system with redundant metadata information only for management tasks. This means that the AIPs must be exportable from the file system even if the repository software fails or can no longer be operated. On the other hand, this possibility should be open for the administrator only. For the archivists there should be only one entry point (the repository itself) to the AIPs, coupled with a user management system. In 2006, none of the repositories inspected was able to meet these requirements. So we decided to build a new one which suits the requirements of an archive. Needless to say,

“archive” here means the traditional memory institution. If you decide to build a new repository on your own, the headache is even growing. Is it possible to construct a repository for all kinds of digital objects or should there be one repository for each type? Presumably, many archivists and librarians would opt for the “one fits all” solution. But in practice, differences can be noted: At the moment, some archives concentrate entirely on only one object type. They’re working on a fully automated import function and a suitable repository. Other objects are expected to come into the same repository later, but the practical plannings for this are postponed.

This is a common strategy of traditional archives: They are looking for solutions for current digital records. Of course, it is important to save this information. Concentrating on one type of objects is also a way to reduce complexity. But at the same time other potentially important records like e.g. databases are neglected. Therefore, we’ve decided to build a repository which from the very first day can import all types of digital records. A metadata model, which covers about three dozens of core metadata, stands in the heart of this repository. Dataset-ID and file-ID, signature, title, description, provenance, time, state, creator and others are collected in a structured way. Many of these can be captured automatically. For each record type, other structured data can be implemented. In the case of databases, there are fields for the number of datasets or columns. Furthermore, non-structured metadata or documentation can also be used. Documentation is always welcome, but except for the above mentioned structured metadata we don’t make an effort to fill each logical information unit in its own data field.

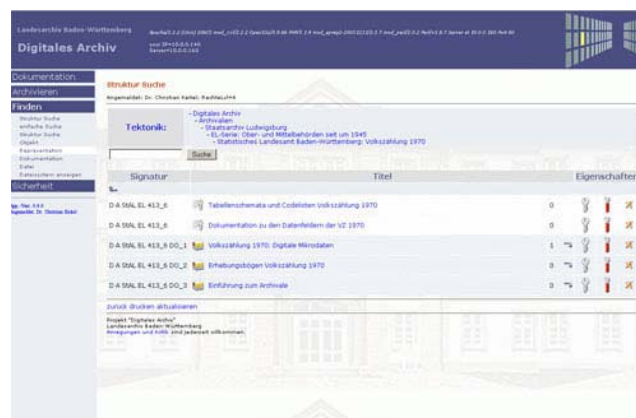
The combination of core metadata, expanded metadata and documentation makes it possible to define only one way of transferring records into the repository. Each object could be sent into the repository manually. But with IngestList it’s possible to transfer them automatically.

Another important feature is the protocol function. It is clear that a repository comprises a lot of duties. Many of these should be listed in a journal so that a future user can consult them in order to verify the trustworthiness. But if each task produces a journal of its own, this would result in a mass of information in a lot of different places. Therefore, we’ve decided to bring all the valuable protocol information together in two kinds of journals: One for each AIP and one for the archive as a whole. These journals aren’t log files, and they are not written in some proprietary file format as it is important to keep them readable for the near and the remote future. For this reason, both protocol types are written as XML files. Each one has its hash value so that it’s difficult to change them without being noticed.

The reduction to a small number of metadata, no more than two protocol types and only one way into the repository helped us to keep the complexity down. As a consequence it was possible for the project team to develop and to implement the digital repository DIMAG with just three persons in 2006.

DIMAG stands for Digitales Magazin (digital storeroom). It is able to hold all kinds of digital objects.

In 2008 the digital repository comprises more than 18.000 digital records, including databases, pictures and textual records. Due to the reduced number of metadata fields and protocol types it is not complex to handle DIMAG.



DIMAG

As previously mentioned, DIMAG is able to handle all types of digital objects; the Landesarchiv Baden-Württemberg keeps nearly all archival objects in it. But there is an exception to every rule. Although it would also be possible to keep websites in DIMAG, these are preserved in BOA (Baden-Württembergisches Online-Archiv). This system is run by the Bibliotheksservice-Zentrum (Library Service Centre) Baden-Württemberg (BSZ), a support institution for libraries and archives. The two state libraries and the Landesarchiv cooperate in the archiving of websites in this system. In this case, complexity was reduced by sharing the risks with other memory institutions on the basis of a common object type; in other words it was reduced by collaboration.

Acknowledgements

Rolf Lang has programmed and implemented IngestList and DIMAG. Kai Naumann has taken the majority of objects into the digital archive. The author thanks these two colleagues for their invaluable contributions. Special thanks also to Heidrun Wiesenmüller for further discussion and suggestions.

Notes

Further information can be found at <http://www.landesarchiv-bw.de>. Use the full text search (entering “DIMAG”) or see under “Fachinformationen” >>>> “Elektronische Unterlagen”.

A Logic-Based Approach to the Formal Specification of Data Formats

Michael Hartle, Arsene Botchak, Daniel Schumann, Max Mühlhäuser

Technische Universität Darmstadt
Telecooperation Group
Hochschulstr. 10
D-64289 Darmstadt, Germany
{nhartle,max}@tk.informatik.tu-darmstadt.de

Abstract

Processing information stored as data in a specific data format is tightly coupled with software implementations that handle necessary elementary processes such as reading and writing. These implementations depend on specific technological environments and thus age due to rapid technological change. The resulting effective loss of information is a major problem for Digital Preservation. In order to provide for persistent, authentic access to stored information, this paper presents a logic-based approach for the formal specification of data formats.

Introduction

What turns data into information is the knowledge on its semantics, its intended meaning. If this knowledge is lost, so is our access to information that is contained in data. A good example from history was the inability to read ancient hieroglyphic Egyptian script for more than a millennium, fortunately solved by the happenstance of the Rosetta Stone. Only by the lucky circumstance of it carrying three distinct translations of a decree, it enabled the inference of the meaning of hieroglyphs in the early 19th century (Solé, Valbelle, and Rendall 2002).

For digital information, the problem of preserving the knowledge of its intended meaning, its data format, is a lot more complex. We do not have a small set of languages like hieroglyphic Egyptian with distinguishable symbols in use, but rather a variety of different data formats on binary data. Each of them defines the meaning of bits and bytes essentially depending on context, so for accessing contained information, establishing the meaning of data from context needs processing. Yet for this processing, we depend on implementations that are expensive to create, do age over time and become obsolete due to rapid technological change.

Research Problem

Our central research problem is that the current state of specifying data format knowledge is based on semi formal, textual specifications. As these documents are intended for human engineers, application of this knowledge to a problem inevitably depends on human labour,

needed for developing suited implementations for a specific technological environment and purpose.

Now, rapid technological change of environments (e.g. hardware, operating systems, programming languages) combined with a variety of processing purposes (e.g. reading, writing, validating, repairing, optimizing) and the ongoing development of data formats constantly retriggers the need for a new development cycle. Complicating matters, reuse is often severely limited, as adaptation of existing source code can be next to impossible due to radical differences in suited implementations. Taking X.509 security certificates as example, developing software can result in widely different implementations for writing them on a Java mobile phone, for reading them in a batch using C++ on a Linux server or for validating them using Assembler on an memory-constrained embedded system.

Developing format-compliant implementations is a highly complex task, yet at the same time, human engineers have cognitive limits and make mistakes. The cost for developing an implementation, e.g. for sufficiently qualified labour, puts economic limits to feasibility for both public institutions and private companies.

Regarding public institutions, current Digital Preservation practices such as evaluating the risk of data format obsolescence in regular intervals and planning for timely data migration tell of this problem. For private companies, there must be a commercial incentive for the development and maintenance of products in support of a specific data format - the monetary value of information contained must match the cost associated with its implementation and support in practice. If the monetary value does not match its cultural or scientific value on a short timescale, products are discontinued or not developed, resulting in a loss of required processing means, the underlying data format knowledge and thus ultimately of access to contained information.

Contribution

For Digital Preservation of information in arbitrary data formats, the current practice of semi-formal, textual specifications and the subsequently required human engineering effort is too expensive to guarantee long

term access to information, not speaking about other usual problems such as format-compliance of implementations and authenticity of data.

We therefore propose the formal description of data formats in order to make data format knowledge machine-processable in the first place and thus enable its automated application in a scalable manner, e.g. for extracting information from formatted data or for generating skeleton source code for implementations.

Towards that purpose, we recently published the concept of *Bitstream Segment Graphs* (BSGs) for describing the composition of data (Hartle et al. 2008a). In this paper, we build upon BSGs and contribute a logic-based approach for formal data format specification.

Related Work

Data formats are not only a subject in Digital Preservation, but rather a cross-cutting concern that appears in other disciplines of research as well:

- In *Multimedia*, motivations for research on data formats were the need to specify data formats for MPEG-4, e.g. for Part 2 (Visual) (ISO 2004) on the one hand and the *Universal Multimedia Access* (UMA) vision (Vetro, Christopoulos, and Ebrahimi 2003) in the context of MPEG-21 (ISO 2007) on the other hand, part of which focuses on content adaptation and filtering. The former led to MSDL-S (Eleftheriadis 1996) and its successor Flavor/XFlavor (Eleftheriadis and Hong 2004), whereas the latter resulted in BSDL (ISO 2008). In this domain, contributions in literature are basically restricted to high-level descriptions of bitstreams.
- Regarding *Telecommunication*, the main motivation was the need to specify an efficient representation of a data model in an interoperable manner. This has led to the Abstract Syntax Notation One (ASN.1) (ITU-T 1997), the generic Encoding Control Notation (ITU-T 2002b) and specific standard encodings such as CER or DER (ITU-T 2002a). For arbitrary data formats that do not fit into these encodings, universal applicability is sometimes claimed for the combination of ASN.1 & ECN, yet such a claim has neither been proven nor substantiated for these two highly complex specifications.

Other disciplines also touch upon the subject of data formats, e.g. the Semantic Web with the problem of making information accessible to machine reasoning, or IT Security with the problem of testing application robustness by the introduction of data errors, so-called *fuzzing* (Miller, Fredriksen, and So 1989)

Approach

In general, we assume a data format to define a lossless digital representation of some structured information for purposes of storage and transmission. A data format therefore defines a set of finite, consecutive bit sequences and a set of structured information. Both sets may be infinite in size and have a one-to-one correspondence.

We thus assume that there exists a bijective mapping function between both sets (for *parsing* and *serialisation*) as well as functions for deciding the membership in either set. For practicability, we require that all three problems (bijective mapping as well as membership in either set) are computable and decidable, that is, there exists a Turing machine that always computes an answer to the problem and halts.

Computational Complexity

Bijectivity of the mapping function does not limit its computational complexity, as it was shown that every single-tape Turing machine can be converted into a logically reversible 3-tape Turing machine (Bennett 1973). Moreover, no general formalism can exist that exactly covers the set of decidable problems, as follows from the *Halting Problem* (Hopcroft and Ullman 1979). Therefore, describing arbitrary data formats requires a formalism which is equal to the Turing machine in computational power. Such a formalism inherits the Halting Problem and thus cannot guarantee decidability by itself.

Decomposing the problem

In order to decompose the problem of formal data format specification, we define a *data format instance* as the bijective mapping between a pair of elements from both sets. We further define a *data format* as a potentially infinite set of data format instances, with the definition intentionally being analogous to that of a formal language (Mateescu and Salomaa 1997).

We therefore decompose the problem of formal data format specification into the problem of describing arbitrary data format instances and the problem of describing a possibly infinite set of bijective mappings.

Model

For the first problem, we have recently proposed a model for describing arbitrary data format instances using the *Bitstream Segment Graph* (Hartle et al. 2008a), which has also been applied for describing exploits in IT Security (Hartle et al. 2008b). For the latter problem, we build upon the BSG model and propose a logic-based approach through fixed-point deduction of BSG instances.

Describing arbitrary data format instances

An abbreviated introduction into Bitstream Segment Graphs is given in this subsection. For a more formalized description, the reader is kindly referred to (Hartle et al 2008a).

Entities

A *bitstream segment* is a finite, consecutive bit sequence such as 01000001. A *bitstream source* is a defined bitstream segment that is to be described and which follows a certain data format, e.g. a specific image file or a network packet.

A *bitstream transformation* is a bijective mapping of input bitstream segments to output bitstream segments, limited to one of the following normalisations:

Bitstream segment type	Used in encoding?	Used in transformation?	Coverage
Generic	no	no (as input)	0
Primitive	yes	no (as input)	1
Structure	no	segmentation (as input)	length-weighted coverage of successors
Transcode	no	transformation (as input)	coverage of successor
Fragment	no	concatenation (as input)	coverage of successor
Composite	no	concatenation (as output)	coverage of successor

Table 1: Bitstream segment types.

- the *segmentation transformation* that splits one input bitstream segment into two or more ordered output bitstream segments (1:n),
- a class of *block transformations* which transform one input bitstream segment into one output bitstream segment (1:1), and
- the *concatenation transformation* that joins two or more ordered input bitstream segments into one output bitstream segment (n:1).

Examples for these normalised bitstream transformations are the segmentation of a data structure into its fields, block transformations such as GZIP compression, AES encryption or Reed-Solomon error-correction, or the aggregation of a fragmented multimedia stream in an Apple QuickTime container. Arbitrary (n:m) bitstream transformations can be constructed through sequential composition of multiple normalised transformations. A bitstream transformation connects input and output segments as *predecessors* and *successors*, respectively. No cycles may be formed through bitstream transformations either directly or indirectly.

A bitstream encoding is a bijective mapping between a bitstream segment and a typed literal, representing some information. For example, the bit sequence 01000001 represents the number 65 for a big-endian unsigned integer encoding, whereas for an ASCII encoding, it represents the letter A.

Every bitstream segment belongs to one of 6 bitstream segment types, depending upon its participation in bitstream transformations and bitstream encodings as listed in Table 1. For example, it may be a *structure* composed from two or more successor bitstream segments, a *primitive* if it represents an encoded literal, or a *generic* if it does not participate in a bitstream transformation or a bitstream encoding.

The *coverage* of a bitstream segment is a measure in the range between 0 and 1 and expresses how completely a bitstream segment is mapped to encoded literals through its successor(s). It is computed depending on the bitstream segment type (see Table 1). For example, for a structure bitstream segment a with two primitive segments as successors, the coverage of a would be 1. In case of one primitive segment and a generic segment of equal length as successors, the coverage of a would be 0.5. The coverage of a BSG instance refers to that of its bitstream source.

A *Bitstream Segment Graph* (BSG) is now a rooted, acyclic graph that is defined from a bitstream source, a set of bitstream transformations and a set of bitstream

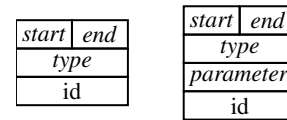


Figure 1: Visual representations: generic, structure and composite bitstream segments (left); fragment, primitive and transcode bitstream segments (right).

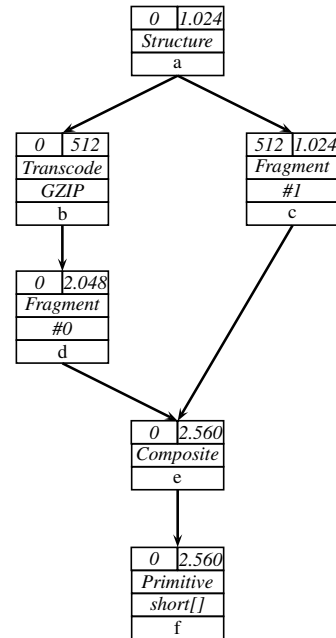


Figure 2: Minimal example of a BSG instance.

encodings, where the nodes correspond to bitstream segments and the edges to transformations. It describes the composition of a bitstream source from primitive bitstream segment(s). For a visual representation of a BSG instance, bitstream segments are depicted as in Figure 1.

Properties

A bitstream segment x has a set of namespaced *properties*, denoted as $ns:property(x, v_0, \dots, v_n)$. For the BSG model, this includes placement information such as an inclusive `bsg:start` position, a `bsg:length` and an exclusive `bsg:end` position, all measured in bits and relative to the context provided by its predecessors. For

Predicate	Behaviour
<code>math:lt(?a,?b)</code>	Tests the formula $?a < ?b$.
<code>math:lte(?a,?b)</code>	Tests the formula $?a \leq ?b$.
<code>math:eq(?a,?b)</code>	Tests the formula $?a = ?b$.
<code>math:product(?a,?b,?c)</code>	Computes the formula $?a * ?b = ?c$ if two parameters are ground and no division by zero occurs, and assigns the result to the third variable parameter. Tests the formula if all parameters are ground.
<code>math:sum(?a,?b,?c)</code>	Computes the formula $?a + ?b = ?c$ if two parameters are ground and assigns the result to the third variable parameter. Tests the formula if all parameters are ground.
<code>util:concat(?a,?b,?c)</code>	Concatenates ground strings $?a$ and $?b$ and binds the result to variable $?c$. Tests whether the concatenation of $?a$ and $?b$ corresponds to $?c$ if all parameters are ground.
<code>util:sourceLength(?a,?b)</code>	Gets the length in bits of the ground file reference $?a$ and binds it to variable $?b$. Tests whether file reference $?a$ has length $?b$ in bits if both are ground.
<code>util:skolem(?a,...,?c)</code>	<i>Skolem function provided for existential quantification.</i> Maps the set of ground parameters $(?a, \dots)$ to a value and binds it to variable $?c$. Maps a ground $?c$ to a set of values and binds them to variables $(?a, \dots)$. Tests whether $(?a, \dots)$ and $?c$ map to each other if all parameters are ground.
<code>util:value(?a,?b)</code>	Decodes the contained literal of a ground primitive bitstream segment $?a$ if it is <code>bsg:resolved</code> , and assigns the result to variable $?b$. Tests whether the bitstream segment $?a$ contains the literal $?b$ if both parameters are ground.

Table 2: List of computable predicates.

example, the first successor segment of a structure segment starts at bit 0. Further properties include the bitstream segment `bsg:type`, one or more `bsg:semantics` as identifiers or a `bsg:codec` identifier for transcode bitstream segments. For example, for the segments b and c in Figure 2, we can state properties such as `bsg:start(b,0)`, `bsg:length(c,512)` or `bsg:codec(b,GZIP)`.

Relations

Between any two bitstream segments x and y , namespace *relations* may exist, denoted as `ns:relation(x,y,v0,...,vn)`. For the BSG model, this includes neighbourhood relations between bitstream segments in a structure bitstream segment as `bsg:leads` and `bsg:follows`, and composition relations such as `bsg:successor` and `bsg:predecessor` with `bsg:firstSuccessor` and `bsg:lastSuccessor` as special cases. For example, for the segments a , b and c in Figure 2, we can state relations such as `bsg:firstSuccessor(a,b)`, `bsg:leads(b,c)` and `bsg:predecessor(c,a)`.

Using suited types of bitstream transformations and encodings, the composition of arbitrary data format instances can be described using BSG instances. Besides the visual representation, we can represent a BSG instance through facts regarding BSG-related properties and relations.

Describing possibly infinite sets of data format instances

We define a potentially infinite set of bijective data format instances through the set of stable models resulting from a set of first-order logic rules, expressed as implica-

tions or biconditionals. For rules, predicates are used that refer to either deduced or computed facts. In terms of existing logic languages, it resembles Datalog (Ullman 1989) extended with functions.

Deducible predicates refer to facts that were either given initially or subsequently deduced through rules. They are not limited to BSG-related properties and relations only, but may also include predicates for intermittent facts which may be needed for deducing a BSG instance. For deduced predicates, the open world assumption applies, as a currently unknown fact may become known later. *Computable predicates* refer to facts that can be computed directly (see Table 2). They handle aspects such as decoding the literal $?l$ of a primitive bitstream segment $?x$ from the so-far deduced, partial BSG instance through `bsg:value(?x,?l)`, or for solving the equation $?v=?u+1$ through `math:sum(?u,1,?v)` if either $?u$ or $?v$ are known. These predicates can choose between the open world assumption and the closed world assumption, as they can decide to refute facts that will always fail, such as `math:sum(1,2,4)`.

Predicates have parameters that can either be *ground* and thus have a specific value, or be a *variable*. A *mode* of a predicate declares for each of its parameters whether it is ground or variable. Computable predicate may support arbitrary modes, e.g. allowing `math:sum` to compute `math:sum(?u,4,5)` as well as `math:sum(1,?v,5)` and `math:sum(1,4,?w)`, or test `math:sum(1,4,5)`. Using these types of predicates, we can build rules as implications or biconditionals. These rules can be partitioned into model-specific rules that capture properties and relations inherited from the BSG model itself, and format-specific rules that represent data format knowledge. For example, a BSG-specific rule is that two

#	Rule
M1	$\text{bsg:source}(?a,?f) \ \& \ \text{util:sourceLength}(?f,?l) \rightarrow \text{bsg:start}(?a,0) \ \& \ \text{bsg:length}(?a,?l)$
M2	$\text{bsg:length}(?a,?l) \ \& \ \text{bsg:end}(?a,?e) \ \& \ \text{math:sum}(?s,?l,?e) \rightarrow \text{bsg:start}(?a,?s)$
M3	$\text{bsg:start}(?a,?s) \ \& \ \text{bsg:end}(?a,?e) \ \& \ \text{math:sum}(?s,?l,?e) \rightarrow \text{bsg:length}(?a,?l)$
M4	$\text{bsg:start}(?a,?s) \ \& \ \text{bsg:length}(?a,?l) \ \& \ \text{math:sum}(?s,?l,?e) \rightarrow \text{bsg:end}(?a,?e)$
M5	$\text{bsg:start}(?a,?s) \ \& \ \text{bsg:length}(?a,?l) \ \& \ \text{bsg:end}(?a,?e) \rightarrow \text{math:sum}(?s,?l,?e)$
M6	$\text{bsg:leads}(?a,?b) \leftrightarrow \text{bsg:follows}(?b,?a)$
M7	$\text{bsg:leads}(?a,?b) \ \& \ \text{bsg:end}(?a,?p) \leftrightarrow \text{bsg:follows}(?b,?a) \ \& \ \text{bsg:start}(?b,?p)$
M8	$\text{bsg:firstSuccessor}(?a,?b) \rightarrow \text{bsg:successor}(?a,?b)$
M9	$\text{bsg:lastSuccessor}(?a,?b) \rightarrow \text{bsg:successor}(?a,?b)$
M10	$\text{bsg:successor}(?a,?b) \rightarrow \text{bsg:predecessor}(?b,?a)$
M11	$\text{bsg:successor}(?a,?b) \ \& \ \text{bsg:leads}(?b,?c) \rightarrow \text{bsg:successor}(?a,?c)$
M12	$\text{bsg:successor}(?a,?b) \ \& \ \text{bsg:follows}(?b,?c) \rightarrow \text{bsg:successor}(?a,?c)$
M13	$\text{bsg:firstSuccessor}(?a,?b) \rightarrow \text{bsg:start}(?b,0)$
M14	$\text{bsg:lastSuccessor}(?a,?b) \ \& \ \text{bsg:length}(?a,?c) \rightarrow \text{bsg:end}(?b,?c)$
M15	$\text{bsg:lastSuccessor}(?a,?b) \ \& \ \text{bsg:end}(?b,?c) \rightarrow \text{bsg:length}(?a,?c)$
M16	$\text{bsg:start}(?a,?s) \ \& \ \text{bsg:length}(?a,?l) \ \& \ \text{bsg:end}(?a,?e) \ \& \ \text{bsg:type}(?a,?t) \ \& \ \text{bsg:source}(?a,?f) \rightarrow \text{bsg:resolved}(?a)$
M17	$\text{bsg:successor}(?a,?b) \ \& \ \text{bsg:start}(?b,?s) \ \& \ \text{bsg:type}(?b,?t) \ \& \ \text{bsg:resolved}(?a) \rightarrow \text{bsg:resolved}(?b)$

Table 3: List of model-specific rules.

neighbouring bitstream segments b and c share a boundary, so from the facts $\text{bsg:follows}(b,c)$ and $\text{bsg:end}(b,512)$, the fact $\text{bsg:start}(c,512)$ can be concluded. From the data format instance in Figure 2, we could assume as format-specific rule that from the facts $\text{bsg:source}(a,...)$ and $\text{bsg:firstSuccessor}(a,b)$, the fact $\text{bsg:type}(b, \text{'bsg:transcode'})$ follows.

For deducing a BSG instance, initial knowledge on a specific bitstream source is given, such as the fact $\text{bsg:source}(a, \text{'oi2n0g16.png'})$. Through a series of iterative steps, the set of rules is applied in a monotone deduction process. In each step for every rule, it is tried to match the antecedents with previously deduced knowledge. If the antecedent of a rule matches, then for its conclusion, the computable predicates are tested and the deducible predicates are deduced. Should a computable predicate fail in this test, the reasoning process aborts, as a conclusion does not hold. This allows the use of validation rules that assert certain properties, e.g. that for all bitstream segments, its respective bsg:start and bsg:length have to sum up to its bsg:end , which can be violated in case of contradictory information contained in a damaged or erroneous bitstream source. When no new facts are deduced in a step, then a fixed point consisting of the deducible facts of a BSG instance is reached.

If a fixed point is reached, the resulting BSG facts can then be translated into a BSG instance for that bitstream source. This requires post-processing steps such as assigning the generic bitstream segment type whenever no type was deduced for a bitstream segment. The deduction of a BSG instance therefore can either

- abort with a computable predicate refuting a fact in a rule conclusion, indicating that a conclusion does not hold and thus the bitstream source does not conform to the specified data format,

- reach a fixed point with a coverage $x < 1$, indicating that there are bitstream segments in this data format instance not specified in the set of rules, or
- reach a fixed point with a coverage $x = 1$, indicating that this data format instance is completely covered by the set of rules.

Building a set of rules as data format knowledge is typically an incremental process. It starts with the collection of bitstream sources for a corpus that represents a specific format, and the definition of an initial set of rules. This set of rules can be improved step-by-step by computing the BSG instance for every bitstream source in the corpus and computing its coverage. One then can select BSG instances with a coverage $x < 1$ and focus on generic bitstream segments which need to be described further through additional rules. Actual knowledge on how these generic bitstream segments are actually composed may come from consulting textual specifications, existing implementations or through try-and-error reverse engineering efforts. Repeating this process increases the overall coverage of BSG instances in the corpus. For a corpus, a *fitting* set of rules is found if the coverage reaches 1 for all of its BSG instances.

Evaluation

In order to apply our approach, we implemented a reasoning system in Java, defined suited interfaces for processing bitstream transformations and bitstream encodings, and implemented components for handling certain transformations and encodings as required.

Setup

For evaluation, we decided to describe a small subset of the Portable Network Graphics (PNG) image format. We required that of this subset, some data format instances

#	Rule
F1	<code>bsg:source(?a,?f) → bsg:semantics(?a,'png:root')</code>
F2	<code>bsg:semantics(?r,'png:root') → util:skolem('F2',?r,?s) & bsg:type(?r,'bsg:structure') & bsg:firstSuccessor(?r,?s) & bsg:semantics(?s,'png:signature')</code>
F3	<code>bsg:semantics(?s,'png:signature') → util:skolem('F3',?s,?f) & bsg:leads(?s,?f) & bsg:semantics(?f,'png:chunk')</code>
F4	<code>bsg:semantics(?c,'png:chunk') → util:skolem('F3',?c,?l) & bsg:firstSuccessor(?c,'png:chunk') & bsg:semantics(?l,'png:chunk-length')</code>
F5	<code>bsg:semantics(?l,'png:chunk-length') → util:skolem('F5',?l,?t) & bsg:leads(?l,?t) & bsg:semantics(?t,'png:chunk-type')</code>
F6	<code>bsg:semantics(?l,'png:chunk-length') & bsg:value(?l,0) & bsg:leads(?l, ?t) & bsg:successor(?ch,?l) → util:skolem('F6',?l,?t,?ch,?cr) & bsg:lastSuccessor(?ch,?cr) & bsg:leads(?t,?cr) & bsg:semantics(?cr,'png:chunk-crc')</code>
F7	<code>bsg:semantics(?l,'png:chunk-length') & bsg:value(?l,?v) & math:lt(0,?v) & bsg:leads(?l,?t) & bsg:successor(?ch,?l) & math:product(?v,8,?lv) → bsg:leads(?t,?d) & bsg:leads(?d,?cr) & bsg:lastSuccessor(?ch,?cr) & bsg:length(?d,?lv) & bsg:semantics(?d,'png:chunk-data') & bsg:semantics(?cr,'png:chunk-crc')</code>
F8	<code>bsg:semantics(?t,'png:signature') → bsg:type(?t,'bsg:primitive') & bsg:encoding(?t,'http://www.dataformats.net/2008/04/bsg-encodings#ascii-string') & bsg:length(?t,64)</code>
F9	<code>bsg:semantics(?t,'png:chunk-length') → bsg:type(?t,'bsg:primitive') & bsg:encoding(?t,'http://www.dataformats.net/2008/04/bsg-encodings#msbf-uint') & bsg:length(?t,32)</code>
F10	<code>bsg:semantics(?t,'png:chunk-type') → bsg:type(?t,'bsg:primitive') & bsg:encoding(?t,'http://www.dataformats.net/2008/04/bsg-encodings#ascii-string') & bsg:length(?t,32)</code>
F11	<code>bsg:semantics(?t,'png:chunk-crc') → bsg:type(?t,'bsg:primitive') & bsg:encoding(?t,'http://www.dataformats.net/2008/04/bsg-encodings#msbf-uint') & bsg:length(?t,32)</code>
F12	<code>bsg:successor(?ch,?t) & bsg:semantics(?ch,'png:chunk') & bsg:semantics(?t,'png:chunk-type') & bsg:value(?t,?v) → util:concat('png:chunk:',?v,?ct) & bsg:semantics(?ch,?ct)</code>
F13	<code>bsg:successor(?r,?c) & bsg:semantics(?c,'png:chunk') & bsg:end(?c,?ce) & bsg:length(?r,?rl) & math:lt(?ce,?rl) → util:skolem('F13',?c,?ce,?r,?rl,?nc) & bsg:leads(?c,?nc) & bsg:semantics(?nc,'png:chunk')</code>
F14	<code>bsg:semantics(?r,?c) & bsg:semantics(?c,'png:chunk') & bsg:end(?c,?ce) & bsg:length(?r,?rl) & math:eq(?ce,?rl) → bsg:lastSuccessor(?r,?c)</code>

Table 4: Excerpt of format-specific rules for a limited PNG subset. Due to length considerations, this list is limited to describing a PNG image down to the level of chunk structures.

should at least be sufficiently complex as to require all three types of normalised bitstream transformations (segmentation transformation, block transformation and concatenating transformation) from the BSG model.

We found a suited subset of PNG images, namely those where compressed image data is stored as separate fragments in so-called IDAT chunks. For building a suited corpus, we examined the PNG Test Suite (van Schaik 1998) with 156 PNG images for compliance testing, including corrupted files and extreme variants, and selected 8 images with filename pattern `oi?????.png`.

Regarding the granularity of description, we allowed primitive bitstream segments to represent arrays of encoded literals. Without this consideration, the decomposition of arrays such as pixel data into individual pixels would have bloated the resulting description of a data format instance without substantial benefit.

We built a fitting set of rules for our corpus, consisting of 17 model-specific rules (see Table 3) and 36 format-specific rules (see Table 4 for an excerpt).

Data format rules

Regarding model-specific rules, we start with rules on placement regarding a bitstream segment. This begins with a rule for deducing `bsg:start` and `bsg:length` from an initially given `bsg:source` (M1). If any two of `bsg:start`, `bsg:length` and `bsg:end` are given for a bitstream segment, the remaining fact can be deduced (M2-M4). Moreover, if all facts are given for a bitstream segment, it can be validated for ensuring consistency (M5). Further rules include aspects of bitstream segments in structures, such as neighbourhood (M6, M7), successorship (M8-M12), placement (M13-M15) and resolvability (M16, M17), whereas the latter is necessary for decoding the contained literal of primitive bitstream segments.

Finally, we come to format-specific rules on our PNG subset. We start with a rule that deduces the PNG-specific type of 'png:root' for a bitstream source (F1). For such a bitstream segment, we can deduce that there exists a first successor bitstream segment `?s` with

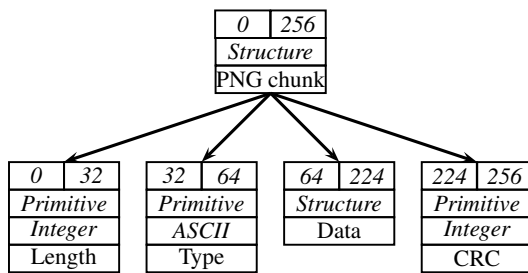


Figure 3: BSG instance for a PNG chunk.

`bsg:semantics(?s, 'png:signature')` (F2). For a 'png:signature', there exists a following 'png:chunk' structure (F3) as shown in Figure 3, which again always begins with a 'png:chunk-length' bitstream segment (F4), followed by a 'png:chunk-type' bitstream segment (F5). If the value of a 'png:chunk-length' is 0, then the 'png:chunk-type' is followed directly by the 'png:chunk-crc' bitstream segment as last successor of the chunk (F6). Otherwise, the 'png:chunk-type' bitstream segment is followed by a variable-length 'png:chunk-data' bitstream segment and again the 'png:chunk-crc' bitstream segment (F7). Details on bitstream segments such as their type, encoding and length are provided for 'png:signature' (F8), 'png:chunk-length' (F9), 'png:chunk-type' (F10) and 'png:chunk-crc' (F11) bitstream segments. The PNG-specific type of the chunk is deduced from the 'png:chunk-type' value and assigned as `bsg:semantics` to the chunk (F12). The remaining rules listed in Table 4 state that if there is space left after a chunk, there exists another one following (F13), otherwise the chunk is the last successor of the bitstream source (F14). Further rules handle chunk-specific aspects, e.g. for the IHDR chunk which contains information on image width and height.

Example deduction steps

For a given initial fact

```
bsg:source('root', 'oi2n0g16.png')
```

the deduction process tries to apply all rules to deduce new facts. In the first step, only the rules F1 and M1 are applicable, which yield the following new facts:

```
bsg:semantics('root', 'png:root') &
bsg:start('root', 0) &
bsg:length('root', 1432)
```

Again, the deduction process tries to apply all rules, this time on an increased set of facts. In step 2, the rules F2 and M4 yield the following:

```
bsg:type('root', 'bsg:structure') &
bsg:firstSuccessor('root', 'scl') &
bsg:semantics('scl', 'png:signature') &
bsg:end('root', 1432)
```

The process of deduction is repeated until either no new facts can be deduced, or a computable predicate refutes a

fact in a conclusion. The resulting facts from the reached fixed point describe a BSG instance for the PNG image `oi2n0g16.png`, which is part of the corpus and has a coverage of 1.0.

Result

After building a fitting set of rules with coverage of 1.0 for our corpus, we tested the set on all remaining PNG images from the PNG Test Suite. We obtained a coverage of 1.0 for 64 images, with the remaining 89 valid images having an average coverage of 0.79. Three corrupt images belonging to the test suite were excluded from the evaluation, as the fitting set of rules did not contain verifying rules for PNG-specific properties.

For a fitting set of rules over the entire PNG Test Suite, additional rules need to be included for palette handling (PLTE and sPLT chunks), transparency (tRNS chunk), background colour (bKGD chunk), textual data (tEXt and zTXt chunks) and other aspects. To estimate the effect of adding further rules, we added two preliminary rules for handling PLTE chunks and re-evaluated our rules on the corpus. We obtained a coverage of 1.0 for 78 images, with the remaining 75 valid images having an average coverage of 0.91.

During evaluation, the deduction process computed a fixed point and halted on all instances. Since errors may be present in a set of rules preventing a fixed point to be reached, a primitive approach on handling the Halting Problem is to place a limit on the iteration steps and abort the deduction beyond that limit. We discovered that the typical number of iterative steps required for our set of rules to reach a fixed point on valid PNG images ranges from 72 up to 170 steps. In case of the image file `oi9n2c16.png`, more than 3,000 iteration steps were required, as compressed image data is present as fragments with a length of 8 bit, each encapsulated into a separate IDAT chunk. This can be considered an extreme example, but demonstrates what is still considered legal in terms of the original specification. Since data format instances of other data formats such as Apple QuickTime movies have a more complex structure which requires an even higher number of iterations, the use of a semi-naive evaluation method for the deduction process as known from Datalog (Ullman 1989) is absolutely essential.

Discussion

The set of rules we tested is quite small, yet describes central elements of PNG files. 'Unexplained' bitstream segments can be readily identified due to the generic bitstream segment type and the coverage measure, and thus allow for incremental development of data format rules. Testing this approach, incrementally adding rules for PLTE chunks to describe palette information had been quite simple and resulted in a significant increase regarding the coverage of nearly all images in the PNG Test Suite.

Regarding data formats in general, our approach maps the diversity of data formats to format-specific data format rules, bitstream transcodings and bitstream encodings. We assume that some bitstream transcodings and a majority of bitstream encodings may be shared among multiple data formats. For example, PNG employs a

scanline transformation to increase the efficiency of a subsequent GZIP compression transformation; the GZIP compression is likely to be reusable, whereas the scanline transformation is highly PNG-specific. The bitstream encodings we encountered so far are basically the ASCII encoding used for PNG chunk types and a bit-endian unsigned integer encoding used for numerical values, which are easily reusable, e.g. in the context of Apple QuickTime.

The set of rules includes model-specific rules that validate the consistency of essential model-specific properties. Due to the complexity of PNG, adding rules for validating all PNG-specific properties is nontrivial and requires specifically corrupted image files for testing the corresponding rules. Our tested set of rules is over-accepting in terms of a formal language when compared to the PNG specification.

We decided to use first-order predicate logic in our approach, yet it may be possible that data formats have rules which are more naturally expressed using fragments of higher-order logics, e.g. when having to express rules on sets of segments. For example, when multiple IDAT chunks are present in an BSG instance, these have to be concatenated in order of their appearance, yet formulating the corresponding rules was non-intuitive. We assume that complex data format rules will at times translate into specialised computable predicates and require larger, more complex sets of rules.

Summary and Conclusion

We have presented an approach for describing arbitrary data formats as a possibly infinite set of data format instances, building upon the Bitstream Segment Graph model. In contrast to previous related work, we can describe arbitrary data format instances down to contained primitives even when real-life aspects such as compression or fragmentation are present. We applied our approach to the description of a sufficiently complex subset of the PNG image format and were able to show that a quite small number of rules is capable of describing a significant part of PNG images. Furthermore, our approach allows the measurement on how completely a set of rules describes a data format instance, which supports the incremental development of data format rules over time.

It therefore provides some means for formal specification of data formats, which may be of use for the specification of new data formats and for the documentation of existing ones. This can especially be helpful for data formats which are undisclosed or which are deviations. For Digital Preservation, a formal data format specification may provide for "a last line of defense" by allowing to extract contained information if a fitting set of rules exists.

Acknowledgements

The authors would like to thank Gina Häußge for feedback, comments and corrections on various drafts of the paper.

References

- Bennett, C. H. 1973. Logical Reversibility of Computation. *IBM Journal of Research and Development* 17(2):525–532.
- Eleftheriadis, A., and Hong, D. 2004. Flavor: A Formal Language for Audio-Visual Object Representation. In *MULTIMEDIA '04: Proceedings of the 12th annual ACM International Conference on Multimedia*, 816–819. New York, NY, USA: ACM Press.
- Eleftheriadis, A. 1996. The Benefits of Using MSDL-S for Syntax Description. Contribution ISO/IEC JTC1/SC29/WG11 MPEG96/M1555.
- Hartle, M.; Möller, F.-D.; Travar, S.; Kröger, B.; and Mühlhäuser, M. 2008a. Using Bitstream Segment Graphs for Complete Data Format Instance Description. In *Proceedings of the Third International Conference on Software and Data Technologies (ICSOF)*.
- Hartle, M.; Schumann, D.; Botchak, A.; Tews, E.; and Mühlhäuser, M. 2008b. Describing Data Format Exploits using Bitstream Segment Graphs. In *Proceedings of The Third International Multi-Conference on Computing in the Global Information Technology (ICCGI)*.
- Hopcroft, J. E., and Ullman, J. D. 1979. *Introduction to Automata Theory, Languages and Computation*. Addison-Wesley.
- ISO. 2004. ISO/IEC 14496-2:2004: Information technology – Coding of audio-visual objects – Part 2: Visual. ISO, Geneva.
- ISO. 2007. ISO/IEC 21000-7:2007: Information technology – Multimedia framework (MPEG-21) – Part 7: Digital Item Adaptation. ISO, Geneva.
- ISO. 2008. ISO/IEC 23001-5:2008: Information technology – MPEG systems technologies – Part 5: Bitstream Syntax Description Language (BSDL). ISO, Geneva.
- ITU-T. 1997. Recommendation X.680 (12/97) — Abstract Syntax Notation One (ASN.1): Specification of Basic Notation. ITU-T, Geneva.
- ITU-T. 2002a. Recommendation X.690 (07/02) — ASN.1 Encoding Rules: Specification of Basic Encoding Rules (BER), Canonical Encoding Rules (CER) and Distinguished Encoding Rules (DER). ITU-T, Geneva.
- ITU-T. 2002b. Recommendation X.692 (03/02) — ASN.1 Encoding Rules: Specification of Encoding Control Notation (ECN). ITU-T, Geneva.
- Mateescu, A., and Salomaa, A. 1997. *Formal Languages: an Introduction and a Synopsis*. Springer Verlag. chapter 1, 1–40.
- Miller, B. P.; Fredriksen, L.; and So, B. 1989. An Empirical Study of the Reliability of UNIX Utilities. Technical report.
- Solé, R.; Valbelle, D.; and Rendall, S. 2002. *The Rosetta Stone: The Story of the Decoding of Hieroglyphics*. Four Walls Eight Windows.
- Ullman, J. D. 1989. *Principles of Database and Knowledge-Base Systems, Volume II*. Computer Science Press.
- van Schaik, W. 1998. PngSuite - The Official Set of PNG Test Images. Available online at <http://www.schaik.com/pngsuite/pngsuite.html>, last accessed 2008-08-015.
- Vetro, A.; Christopoulos, C.; and Ebrahimi, T. 2003. From the guest editors: Universal Media Access. *IEEE Signal Processing Magazine* 20(2):16–16.

National and International Digital Preservation Initiatives

Neil Grindley

Joint Information Systems Committee
Brettenham House (South), 5 Lancaster Place, London, WC2E 7EN
n.grindley@jisc.ac.uk

Abstract

In accordance with the theme of iPres2008, this panel session will consider how 'joined up' (or otherwise) various national digital preservation initiatives are, and whether there is scope to increase levels of co-ordination and collaboration across international boundaries. The panel members will give overviews of some of the activities being undertaken in their respective countries, prefacing a more general discussion involving panel and audience to address issues arising and to identify opportunities and challenges.

Introduction

This panel session is an opportunity to hear descriptions of digital preservation activities relevant to four different countries, and then to discuss those activities in the wider international context. The need for Digital Preservation challenges to be tackled collectively - perhaps even globally - is frequently acknowledged and the iPres conference is a great opportunity to tackle these challenges in a joined-up way. This year's conference theme obviously lends particular relevance to this approach.

Participants in a session such as this could potentially have been chosen from a wide variety of organisations operating in a number of different countries and it should be understood that the chosen panel is meant to represent one example configuration of a group capable of discussing national and international initiatives. Other voices representing other countries and activities beyond the scope of those represented by the panel will of course be critical in validating the discussion, and delegates are invited and encouraged to voice ideas and opinions during the discussion part of the session.

Session Objectives

The scope of this session is potentially enormous and as such, it is acknowledged that it is unlikely to provide anything other than a jumping off point for further discussion. There are, however, multiple imperatives acting upon institutions, governments, funding bodies, etc. to ensure that they are as aware as possible of the range of digital preservation initiatives that are being undertaken around the world. Besides from the obvious benefit of potentially avoiding duplication in terms of effort and resources, there is also a need to ensure that the solutions

and strategies that are developed to tackle problems related to Digital Preservation are the most sustainable, the most widely validated, and of most relevance to a world whose social, cultural and scientific record is increasingly being recorded only in digital form, using methods that are ubiquitous and reliant on content that is globally distributed.

By starting with four brief presentations that describe activities at the national level in four different countries, it is anticipated that indications of strategic overlap will begin to emerge, which may then suggest areas where an increased focus on joint international working may be appropriate.

Contributors

Neil Grindley – The Joint Information Systems Committee (JISC) UK

The JISC is an organization that supports the innovative use of information and communications technology to support teaching and research in UK Higher and Further Education. It achieves this by: careful and selective funding of relevant programmes of work that are firstly endorsed and then overseen by representatives of the community that it serves; and by the formation of strategic partnerships and collaborations with a wide range of organisations both within and beyond the UK. The Digital Preservation and Records Management Programme is the current iteration of a long-running commitment on the part of the JISC to support institutions in keeping digital materials viable and accessible during their entire life-cycle, and incorporates a number of projects, studies and collaborative initiatives that together form the substance of the programme and deliver its benefit to the community.

A brief overview will be given of the more significant pieces of work that are ongoing and indications will be given for emerging funding priorities. This will be considered in the context of other work that is being undertaken in the UK, not only by other programmes supported by JISC, but also by a number of other major agencies and organisations who are interested and involved in Digital Preservation.

Martha Anderson – Library of Congress (USA) – National Digital Information Infrastructure and Preservation Programme (NDIIP)

Martha Anderson will introduce the Library of Congress Programme that has established a network of partners dedicated to collecting and preserving important born-digital information. This national network currently has more than 100 participating entities and has worked on: developing roles and functions for the stewardship of at-risk content; building communities of practice; developing shared services; and building capacity for digital preservation work. The stakeholders in this network represent public and private sector organisations including government agencies, commercial content producers, libraries, archives and technology entities

Steve Knight - The National Library of New Zealand

The National Library of New Zealand (Te Puna Mātauranga o Aotearoa) Act 2003 requires the National Library to collect, preserve, protect and make accessible digital collections, along with traditional paper collections, in ways that ensure current and future access to New Zealand's documentary heritage.

The National Digital Heritage Archive (NDHA) Programme was established in July 2004. Due to be completed in 2009, the NDHA is being developed and implemented in partnership with Ex Libris Group and Sun Microsystems as a commercially viable solution addressing the ingest, workflow, provenance, integrity/authenticity etc issues of institutional digital preservation.

This presentation will outline the National Library's work on digital preservation, how the NDHA fits into that work, what is expected to be delivered through the NDHA, how the organisation is preparing to integrate the new systems, and some comment on the issues going forwards.

Natascha Schumann – The German National Library

The German National Library is one of seven project partners participating in Nestor: The German Network of Expertise in Digital Long-Term Preservation. Natascha Schumann will give an overview of the benefits and the problems of cooperation among diverse communities. She will consider which synergies can be leveraged and which differences have to be taken into account. Issues such as different legal mechanisms for undertaking preservation and different terminologies across communities will be considered. Acknowledgement will also be made of the variances in priorities across disciplines for different approaches to preservation, e.g. the need for 'raw data' curation procedures within the scientific community.

The Panel Discussion

For the concluding discussion, two further panelists will be invited to join the above participants.

Horst Forster - Director, Directorate General for Information and Media - European Commission

Horst Forster will be giving the keynote address at the beginning of the second day of the conference and will be able to speak about initiatives supported by the commission and its potential role in supporting future international collaborations.

Frances Boyle – Director of the Digital Preservation Coalition (DPC)

The Digital Preservation Coalition (DPC) was established in 2001 to foster joint action to address the urgent challenges of securing the preservation of digital resources in the UK and to work with others internationally to secure the global digital memory and knowledge base. The member organisations that constitute the coalition are cross-sectoral in nature and include national libraries and archives, universities, societies and governmental organisations.

Digital preservation activities across communities – benefits and problems

Natascha Schumann

German National Library
Adickesallee 1
60322 Frankfurt am Main, Germany
n.schumann@d-nb.de

Abstract

The problem of digital preservation is not limited to special communities. It concerns all institutions that are involved in the preservation of our cultural heritage. The German network of expertise in long-term preservation of digital resources – nestor brings together different communities to work on solutions for digital preservation. Museums, archives, libraries and scientific institutions are collaborating to create a durable infrastructure which focuses on a wide variety of skills in the area of digital preservation.

This presentation will give an overview of the benefits and the problems of cooperation among diverse communities. Which synergies can be leveraged and which differences have to be taken into account?

The nestor lessons learned show structural differences between the communities as well as common approaches and strategies concerning digital preservation.

Different cooperation models will be presented, as well as legal aspects.

Introduction

The German Network of Expertise in digital preservation is a collaborative project with seven partners. Involved are: The German National Library, the State and University Library Goettingen, the Bavarian State Library Munich, the Computer and Media Service of Humboldt University Berlin, the Institute for Museum Research Berlin, the Federal State Archives Koblenz and the University Hagen (FernUniversität).

Aside from these partners, many other people from different kinds of institutions are engaged in various nestor working groups. Although there are different priorities in relation to digital preservation within the communities, it is obvious that they have to solve a common challenge. Not only the communities represented in nestor are different but there are also differences within each sector, i.e. public and academic libraries, state archives and media archives etc.

Collaborative Working

In order to deal with the challenge of digital preservation, nestor is not only divided in several work packages (WP),

but has also established topic oriented working groups (WG). There are ten working packages in nestor and for each of them one or two of the seven nestor partners are responsible. Some of the WPs were built by reason of content, others are dealing with formal topics.

Project structure: Work Packages

WP 1 is engaged in the distribution of information via the nestor homepage. Three partners share the work on 1. the website itself, 2. the databases and the 3. internal communication (wiki).

The goal of **WP 2** is to promote activities within the archive and museum communities. Therefore special events and workshops are held.

WP 3 is engaged in national and international standardisation activities. The tasks are the development of requirements regarding standardisation, the representation of interests in different standardisation boards and the creation of recommendations. One result of WP 3 is the establishment of a subcommittee for the needs of long-term preservation within the DIN (German Institute for Standardisation). There are two main topics: the first is audit and certification of digital repositories and the second is the standardisation of the ingest process. WP 3 also provides catalogues of criteria for trusted repositories as well as for persistent identifiers.

Within **WP 4**, the activities of the nestor working groups are centralised. More on this later.

Further important parts are training and qualification, which are organised by **WP 5**. This includes the preparation and realisation of training events such as the summer/winter schools, partly organised in collaboration with DPE and DELOS. These are offered twice a year to practitioners as well as to students and others who are interested in learning about digital preservation in general. WP 5 is editor of the "nestor Handbook – An Encyclopaedia in digital Preservation". The authors are experts from different communities and institutions. It is freely available on the website and can be downloaded as a whole or in single files. An important key activity is the development of e-learning modules on digital preservation in collaboration with university partners.

WP 6 is dedicated to international activities concerning digital preservation. The goal is the creation and maintenance of contact to other persons, projects and networks in this field. Further down the line, a European infrastructure should be built and the collaboration i.e. in the field of education extended.

All concerns of PR are clustered in **WP 7**, which is supported by a PR agency.

The evolution of a sustainable infrastructure for digital preservation in Germany is the task of **WP 8**. For the near future, the goal is to continue the work of nestor in a durable organisational form when the current round of funding by the German Ministry of Research and Education expires. Therefore it is necessary to conclude cooperation agreements. A future nestor-organisation should support the institutions tasked with digital preservation, process the information, propose research projects, improve initial and further training and assume other coordination tasks. The goal has to be a permanent and durable organisation and it has to be located at a federal, state and local level.

WP 9 supervises the publication of four studies on different aspects of digital preservation. These studies focuses on raw data, Grid technology and multi-media objects, all in correlation to digital preservation. Additionally the WP wants to initialise a roadmapping process in order to combine ltp-infrastructure with Grid technology.

The project management is the task of **WP 10**.

Project structure: Working Groups

As mentioned before, there are additionally some Working Groups (WG) in nestor. Within these WGs nestor partners collaborate with persons from all institutions dealing with digital preservation. The leadership of the WGs is assumed by two of the project partners, other institutions are invited to collaborate.

The working group **Trusted Repository Certification** works on identifying relevant features and ranges to evaluate existing and emerging digital object repositories in order to form a web of trustworthiness. Those digital repositories can then function as long-term digital archives within various environments. The nestor working group consists of representatives from libraries, archives, museums, data centers, national libraries (Germany, Austria), publishers and certification experts. The working group has developed a catalogue of criteria for trusted digital repositories. Version 1 is published and available on the website, an updated version will be published soon.

The working group **Media** is aspiring to become a centre of knowledge and expertise on best-practice approaches to the problem of long-term accessibility of digital, non-text based media. With the participation of renowned experts on the topic, a virtual meeting point has been established and a handbook on long-term archiving of non-text based media will be published with special consideration of problems regarding file formats, hardware for the creation of archival backup copies and workflow.

One of the goals of the working group **Long-term Preservation Standards** is to achieve interoperability and trustworthiness. Guidelines for the ingest process will be published at the end of the year. The working group is engaged in national and international standardisation boards.

The working group **Grid/eScience and long-term preservation** focuses on synergies between grid computing and long-term preservation. eScience is based on managing tremendous data volumes with Grid technology. The technical dynamic generates a special need for long-term preservation. On the other hand this technology has a potential for the implementation of long-term archive systems. The task of the working group is to outline this new area with its opportunities and risks and to generate a roadmap for the development of long-term preservation.

The working group **Cooperative long-term preservation** promotes a co-operative approach and strengthens binding legal deposit directives. Furthermore, different types of co-operations have been evaluated and the results will be published in 2008. Based on this upcoming study, some topics of this paper will be presented later on. Technical and legal aspects as well as workflow issues related to cooperative long-term preservation are the main topics of the working group. A sub-working group works on recommendations concerning copyright act regarding long-term archiving.

Overview Work Packages

WP 1 – Maintenance of information and communication platform
WP 2 – Digital preservation for Archives and Museums
WP 3 – National and international standardisation
WP 4 – Working Groups <ul style="list-style-type: none"> 4.1 WG on trusted Repository Certification 4.2 WG Media 4.3 WG long-term Preservation Standards 4.4 WG Grid/eScience and long-term Preservation 4.5 WG Cooperative long-term Preservation
WP 5 – Education and Training
WP 6 – International Networking
WP 7 – Public Relations
WP 8 – Sustainable Organisation
WP 9 – Publication of studies
WP 10 – Project coordination

Types of cooperation

There are different kinds of cooperation in the field of digital preservation. As mentioned before, the WG

“Cooperative long-term preservation” have evaluated some existing cooperation projects in Germany.

Types of Archives

On the basis of the OAIS (Open Archival Information System) we can distinguish “independent archives”, “cooperating archives”, “federated archives” and “archives with shared functional areas”, whereas archive is used as a term within this model.

An “independent archive” relates to a single community and may choose its tools and classification systems by itself. This means that it need not conform to standards regarding formats, interfaces etc.

By contrast, “Cooperating archives” have agreements about the use of common standards.

“Federated archives” do not exchange their collections but uses common finding aids. There are two variants, the first one operates with one catalogue in a distributed system and in the other variant the finding aid as well as the object is located at the single archive. Requests were bundled and then sent to the archives.

“Archives with shared functional areas” have agreed to share functional areas and infrastructure.

Chances and Risks

Apart from this model the evaluated institutions state, that cooperation is not the one and only solution, but they consider, that they have to calculate the risks and chances of cooperation.

In the following, some recommendations resulting from the evaluation are listed. They reflect the experiences made by those interviewed and could be regarded as general guidelines.

Even if the cooperation is planned for a limited duration it is helpful to determine long term goals. These goals as well as their realisation should be checked periodically by the partners.

Chances and benefits of cooperation want each partner to benefit from this arrangement. The costs and risks could be shared while efficiency increases by a division of labour. Another important aspect is that collaborative work on special issues brings forward the development within the community.

Some aspects are seen as problematic, i.e. that only one partner might take advantage from the cooperation or that risks and costs are unequally distributed. Another problem mentioned is the danger that extensive approval processes will stifle productivity and innovation.

To prevent these disparities, it is advisable to make these apprehensions explicit and discuss possible strategies before signing an agreement.

There also different assessments regarding the project planning and how detailed it should be. On the one hand, a strongly regulated agreement can avoid misunderstandings. On the other hand, too much planning may lead to over-regulation and slow development.

However, potential conflicts should be discovered early on in the process of cooperation. If possible, all opportunities to deal with them should be taken.

It may also be helpful in the run-up to the cooperation to think about opportunities to terminate the cooperation during the term. First of all, a binding arrangement has to be agreed on, so that in every conceivable scenario, the disposition or the delivery of the data is regulated and assured.

The long-term preservation of digital objects has to be assured in every case. Depending on the form of archive/cooperation, standardised interfaces are necessary, particularly in case of independent archives. Furthermore, it is necessary to agree on common exchange formats.

All in all, as requirements for best practices are trustful relationships as well as an obligatory financial basis.

Further aspects are different legal mandates not only for different communities but also within particular areas. They will regulate what, for how long and in which manner objects will have to be stored. This means that there also might be a transfer from one to another archive after a certain period. Agreements on standards concerning formats, processes etc. are necessary to deal with this.

Conclusions

Despite some difficulties which are not only limited to cooperation in the area of digital preservation, the experience gained within nestor as well as in other cooperation models shows that especially in this field collaborative approaches are very useful and necessary. Even if there are very different initial situations for the communities involved, every institution dealing with the preservation of the digital cultural heritage has to be engaged in digital long-term preservation. Just these different approaches and strategies due to the requirements of the particular sector provide new and innovative perspectives. It is also the variety of strategies as well as research on a European and international level that will generate solutions to meet the challenges of digital preservation. nestor as a network of expertise from different communities shows that cross-sectoral cooperation is a viable approach and that it will build a basis for a future alliance dedicated to preserve our cultural heritage.

References

<http://www.langzeitarchivierung.de/>

<http://www.langzeitarchivierung.de/index.php?newlang=eng>

International Approaches to Web-Archiving Panel Discussion

Thorsteinn Hallgrímsson

National and University Library of Iceland, Arngrímsgata 3, 107 Reykjavík

thh@bok.hi.is

considered for all aspects of web archiving and preservation efforts.

Preserving the World Wide Web

The Web is a separate medium just like books, newspaper, periodicals, CD's, movies et cetera. It is totally digital but the contents are both born digital material and digitized versions of other media. It contains an enormous amount of data, measured in billions of documents. It is also very volatile and it was soon discovered that a lot of its contents is short-lived and disappears. Preserving this medium for the future therefore presents many new problems and challenges to those who attempt it. The first efforts to preserve the Web by archiving web pages were made in 1996 by Australia, Internet Archive and Sweden and by 2000 several efforts were being made. In 2003 most of the institutions that were seriously thinking about preserving the Web, i.e. 11 national libraries and IA, established the IIPC (International Internet Preservation Consortium) and now it has 38 members including 28 national libraries.

To those involved in preserving the Web it is obvious that currently, and increasingly in the future, a large and significant part of our culture will only exist on the Web and therefore this medium must be preserved for the same reasons that most countries strive to preserve and provide access to their cultural and intellectual heritage by collecting it and storing in museums, archives and libraries. If this is not addressed now an important part of our culture, together with most documentation of the cultural change involved, will be lost. Considering that IFLA has more than a hundred national libraries it is valid to ask why only minority has actively started to preserve their national Web space? There is not a simple answer to this but the following one or more of the following reasons certainly play a role:

- Archiving and preserving the web is on the borderline between the library and the information technology (IT) professions, and the methods used reflect that. It is a library collection, but it requires substantial involvement of IT resources for implementing technical solutions and because of the huge volume of documents involved. Library systems have to cope with up to a 100 million records, a web archive must cope with several billion records. This difference in scale must be

- Legal issues and policies are important while in many countries the legal framework for archiving web pages does not exist or is considered to be an obstacle. In recent survey of IIPC members only five countries responded that they had a Law enacted or passed allowing them to collect web pages and archive them. Another four expect a law to be enacted and five are lobbying for a law. Obviously this complicates the issue while national libraries have traditionally relied on the legal deposit law of each country to economically and comprehensively collect and preserved manuscripts and published printed material, and as publishing technology has progressed the libraries extended their collection activity by including physical electronic media like CD-ROM's and some electronic publications like electronic journals. If the traditional axiom of the legal deposit laws and other collection activity holds true it is therefore an absolute necessity to extend this concept to the web. Still this situation has not prevented many countries from actively working on preserving at least parts of their national web domain by collecting web pages
- National libraries do not agree on what preserving the national web domain really means and what kind of collection building rules should be applied. The web is very different from other media while everybody can input documents without any editing or screening in almost any format one can think of. The contents reflect almost every aspect of the daily life, concerns and issues in those parts of the world that have easy access to the Web and range from the trivial to very important data about society. It must be noted here that in many countries common people do neither have easy access to PC's nor to the Web. Some national libraries have decided to use traditional librarian values, where „quality“ web sites are selected, harvested and catalogued by librarians, and access is by structured search. Others have decided to endeavour to use bulk harvesting to take periodic snapshots of their countries' entire web domain trying to preserve everything with the aid of computer technology. There are several reasons for this. One of which is

the difficulty in establishing what will be of value to future researchers and what will not. To make selections from millions of web sites requires enormous personnel efforts at high costs, whereas costs for data memory storage are decreasing at a rapid rate.

A good solution may be to combine periodic snapshots of the entire national domain with selective collections using thematic/event based criteria, and in some cases select web sites that change very frequently like the news media.

Building and sustaining a web archive incorporates the same main activities as in building a traditional library or archive collection, i.e. selection, collection, registration, access, and preservation. From the outset it is important to realize that because of how volatile the web is, it is practically impossible to collect every object present in the web sites or web domains selected. The data that is published on the Web will not be sent to the libraries for preservation but must be actively and systematically collected (harvested) by the libraries. Therefore preservation of the Web starts with the harvesting activity.

The presentation by the National Library of Australia, Bibliothèque nationale de France, British library, The National and University Library of Iceland and the Danish Netarkivet should give the audience a feeling for what is happening worldwide in trying to preserve the Web.

Training and Curriculum Development Panel Discussion

Frances Boyle

Digital Preservation Coalition
DPC, Innovation Centre, York Science Park, York YO10 5DG, UK.
fb@dpconline.org

Abstract

The Training and Curriculum Development Session at iPRES 2008 will bring together a range of experts actively involved in the international digital preservation training arena. Through a series of presentations and modulated discussion it is hoped that the session will be of interest and value to both practitioners and managers who are seeking to find a path through this sometimes fragmented area.

Aims of the Session

The session will be as interactive and participative as possible in what is hoped will be a stimulating and timely topic. The session is a mix of case studies from the panel, followed by an interactive discussion between the audience and the speakers. Our hope is that we can move away from simple reportage of the contributors' projects to a session which engages with the audience. The overall objective is that there will be clear outcomes and outputs from the dialogue which would begin to meet the identified challenges and take forward the digital preservation training agenda. These will be made available to the community after the event.

Structure of the Session

The contributors to the session bring a host of international experience and expertise with them. They will each discuss their own work and training programmes. They will share what they see are the key challenges for the community, how they aim to address these and what the future holds for their initiatives and training programmes. During the interactive session the audience will be invited to identify what their key challenges are and how these map onto those identified by the speakers. For example are there common shared issues globally or are there concerns which are particular to specific regions/sectors etc? Are there differences between the training providers and the community? We would hope to be able to gather concerns, needs and requirements from the audience and synthesise these into challenges, practitioner wishes and suggested solutions which would inform all the training programmes.

The session will be chaired by Frances Boyle from the Digital Preservation Coalition (DPC).

Contributors

Nancy McGovern - Interuniversity Consortium for Political and Social Research (ICPSR)

The Digital Preservation Management Workshop, created by Anne Kenney and Nancy McGovern at Cornell University is celebrating its fifth anniversary this year.

Beginning in April, 2008, the National Endowment for the Humanities has provided funding to ICPSR to continue and expand the workshop, transferring its home from Cornell to the University of Michigan.

Nancy McGovern will look at the impact of the workshop, assess its successes and areas that need greater development, and outline the plans and goals for its next phase.

Kevin Ashley - University of London Computer Centre (ULCC)

The University of London Computer Centre (ULCC) has been involved with the Digital Preservation Training Programme (DPTP) since 2005. The programme is predicated on the need for institutions to combine organisational and technological perspectives to devise an appropriate response to the challenges that digital preservation requirements present. It is aimed at managers in institutions who are grappling with fundamental DP issues. The programme began life as a project funded by JISC under its Digital Preservation and Asset Management programme. That project was led by ULCC working with its partners the Digital Preservation Coalition (DPC) and Cornell University.

We will learn from Kevin Ashley how the DPTP had developed in the UK and will also gain insight into some related work undertaken in Australia.

Rachel Frick - The Institute of Museum and Library Services (IMLS)

The Institute of Museum and Library Services (IMLS) is the largest US cultural funding agency and primary source of federal support in the US. This year IMLS distributed

over \$20.3 million through its 21st century library professional program; a large portion was allocated to programs that provide curricular support and training in the area of digital preservation.

Rachel will share with the audience her insights and perspective into the role of a funding agency. She will look at how the IMLS is supporting a number of exciting activities in training and curriculum development in digital preservation.

Joy Davidson -The Digital Curation Centre (DCC)

The DCC works with many international organizations and is a leading player in digital curation and the management of research data outputs. They are involved in large scale pan European digital preservation projects including DPE, CASPAR and Planets. All of these projects have an associated outreach and training programme

Joy will share with us her recent work with various international bodies and projects to promote a collaborative approach to training and to reduce duplication of effort in this area.

References

DPTP - <http://www.ulcc.ac.uk/dptp>

ICPSR- <http://www.icpsr.umich.edu/ICPSR/>

IMLS - <http://www.ims.gov/>

DCC - <http://www.dcc.ac.uk>

Author Index

Abrams, Stephen	86	Ferreira, Miguel	235	Lee, Christopher	13
Addis, Matthew	134	Fischer, Randall	223	LeFurgy, William	140
Ashley, Kevin	70	Fitzgerald, Brian	140	Lunghi, Maurizio	242
Ball, Alexander	107	Fugazza, Cristiano	242	McDonald, Robert	145, 197, 250
Barbedo, Francisco	235	Greenberg, Saul	32	McLeod, Rory	127
Beagrie, Neil	1	Grindley, Neil	300	Miller, Ant	134
Becker, Christoph	115	Guttenbrunner, Mark	115	Minkus, Michael	20
Beinert, Tobias	20	Guy, Marieke	70	Minor, David	250
Bellini, Emanuele	242	Hagel, Harald	20	Morrissey, Sheila	86
Besek, June	140	Hallgrimsson, Thorsteinn	305	Mossink, Wilma	140
Borghoff, Uwe	20, 205	Han, Yan	281	Müller, Lars	292
Botchak, Arsene	292	Harmsen, Henk	220	Muir, Adrienne	140
Boudrez, Filip	270	Hatcher, Jordan	70	Murray, Ronald	163
Boyle, Frances	307	Hemmje, Matthias	189	O'Steen, Ben	169
Brocks, Holger	189	Hitchcock, Steve	169	Patel, Manjula	107
Brown, Adrian	169	Houtman, Frank	264	Pennock, Maureen	96
Carpendale, Sheelagh	32	Hunter, Jane	181	Pinsent, Ed	70
Carr, Leslie	169	Hutt, Arwen	145	Rajecki, Keith	175
Castro, Rui	235	Jefferies, Neil	169	Ramalho, José Carlos	235
Chan, Chi Pak	281	John, Jeremy	48	Ras, Marcel	264
Chou, Carol	223	Jones, Sarah	213	Rätzke, Björn	205
Cirinnà, Chiara	242	Jordan, Chris	250	Rauber, Andreas	115
Corujo Luis	235	Kehoe, William	56	Rettberg, Najla	1
Cramer, Tom	86	Kehrberg, Carmen	115	Rieger, Oya	56
Daigle, Bradley	40	Keitel, Christian	287	Rödiger, Peter	20, 205
Dappert, Angela	5, 151	Kelly, Brian	70	Rosenthal, David	274
Damiani, Ernesto	242	Klas, Claus-Peter	189	Ross, Seamus	213, 257
Davis, Richard	70	Klett, Fanny	229	Rusbridge, Adam	257
Ding, Lian	107	Knight, Gareth	96	Ruusalepp, Raivo	213
Dobratz, Suzanne	205	Kosovic, Douglas	181	Saul, Christian	229
Enders, Markus	151	Kozbial, Ardys	145, 250	Schoger, Astrid	20, 205
Faria, Luis	235	Lang, Suzanne	20	Schrimpf, Sabine	28
Farquhar, Adam	5	Lazzarino, Franco	223	Schumann, David	292

Schumann, Natascha	302
Sierman, Barbara	13
Spencer, Amanda	78
Steinke, Thomas	159
Sutton, Don	145
Tarrant, David	169
Van der Hoeven, Jeffrey	93
Van Diessen, Raymond	13
Van Wijngaarden, Hilde	264
Von Suchodoletz, Dirk	93
Walters, Tyler	197
Westbrook, Brad	145
Weston, Christopher	140
Wheatley, Paul	122
Williams, Peter	1
Woods, Kam	62
Wright, Richard	134
Au Yeung, Tim	32